Longer version of a paper that appears in *AAAI Fall Symposium on Knowledge and Action at Social and Organizational Levels*, November 1991

Social and Psychological Commitments in Multiagent Systems

Munindar P. Singh*

DFKI
(German Research Center for AI)
Postfach 2080
D-6750 Kaiserslautern
Germany

and

Center for Cognitive Science
and Dept of Computer Sciences
University of Texas
Austin, TX 78712
USA

msingh@cs.utexas.edu
msingh@dfki.uni-kl.de

14 June, 1991

abstract

Commitment is a central concept in Artificial Intelligence (AI). At least two kinds of commitment can be identified that have been used in AI—the internal or *psychological* and the external or *social*. While the former has been made explicit in AI theory, the latter has often been ignored (though it is given more importance in other disciplines, and indeed even in AI practice). The many roles of social commitment in AI are discussed. It is argued that while social and psychological commitments are related, they must not be conflated with each other. In particular, thinking directly in terms of social commitments helps us avoid the infelicities of traditional theories of group action and intention. In the full paper, a formalization of these concepts is also proposed that captures their desired properties and interrelationships.

*This research was supported by DFKI.

1 Introduction

At least two kinds of commitment can be identified in work in Artificial Intelligence (AI). One is the familiar notion from Ethics and Distributed Computing where an agent is deemed to make a commitment to others to do certain actions (including actions such as issuing permissions) or to prevent certain conditions. The agent who makes such commitments is socially liable for not acting up to them. Another, quite different, sense of commitment has also been drawing attention recently. This is the kind of commitment that Bratman [1, ch. 2] and Harman [7, p. 94] invoke when they say that an agent is "committed" to his intentions. This is also the notion used in [2, p. 217] and studied in [12]. As I will explain shortly, this sense of commitment also arises under the guise of the "epistemic entrenchment" of an agent's beliefs as in [4] and [7]. I call the first kind of commitment *social* commitment or *S-commitment* and the second kind *psychological* commitment or *P-commitment*. These concepts are the subject of this paper.

The question: *can a group be the locus of believing and intending?* may be raised. The underlying assumption of this paper is that it can unequivocally be answered with a "Yes!" This seems justified by our commonsense intuitions about human and artificial multiagent systems. Groups such as teams, armies, nations and corporations may all be (and are in fact) profitably treated as being single, though potentially complex, agents. The notion of agency is intimately connected to ascriptions of belief and intention—if something can be considered an agent, it can and must be ascribed at least some relevant beliefs and intentions. That such ascriptions are legitimate has also been argued by McCarthy [9] and Dennett [3]: one only has to realize that groups are, at worst, artificial and complex physical systems. One can do better, however, since the "mental properties" of groups depend on their members and on how they are structured. Modern theories of intention and belief, e.g., those of Harman, Bratman and Gärdenfors, all assert that believing or intending entails, or at least is intimately connected to, the notion of P-commitment. So the above question may be posed as: *can a group be the locus of P-commitment?*, and still be answered with a "Yes!"

S-commitments arise when agents interact with other agents. They can be as complex as the social structures in which they arise. Just as P-commitments are needed in understanding simple agents, S-commitments are needed in understanding multiagent systems. Traditional formal theories give preeminence to P-commitment and do not seem to recognize the complexities of S-commitment that are possible and necessary in real systems—they allow S-commitments implicitly, but only in groups which are internally homogeneous, and even then are not quite correct. Indeed, if current theories are right, group action and intention are impossible! Note that since groups are first-class agents, they too can have S-commitments to other agents, who might themselves be groups.

The ideas of commitment (especially S-commitment) as explored in this paper underlie much research into distributed AI, though they are not usually formalized in the way proposed here; e.g., agents collaborating in design have to commit to various issues to arrive at a consistent solution. The motivations for commitments as defined here is both descriptive and prescriptive. Firstly, commitments of both kinds are held by the intelligent systems that we encounter in real-life: humans, corporations and what have you. Secondly, commitments of both kinds can be motivated on grounds of individual and group rationality for limited systems and thus may be prescribed to them or incorporated in their design. Also, the problem of determining how commitments may be formalized so that their important properties can be captured is important—such formalizations usually involve complicated notions such as mutual belief and are, as I explain below, *a fortiori* unimplementable in real-life systems where communication is slow and unreliable. The approach proposed here is not only descriptively valid, but also prescriptively feasible.

As an explanation of my terminology, note that I use *psychological* to refer to anything within an agent, even if that agent is a group of possibly dumb agents, e.g., a termite colony in which

the individual termites do not reflect on the colony's actions and do not have explicit symbolic representations of them. I use *social* for anything arising between agents. Thus the title of this paper could have been "inter- and intra-agent commitments . . . ," with the proviso that intra-agent commitments of a group might rely on inter-agent commitments of its members.

2 Commitments

It should be clear that whatever their interrelationships, P-commitments and S-commitments are, as concepts, quite different from each other:

1. An agent may be psychologically committed to a belief or an intention and yet not be S-committed to it.

2. An agent might be officially S-committed to an intention or belief, but not be psychologically committed to it; e.g., a bad cop might not intend to catch a robber friend of his.

3. An agent, e.g., Robinson Crusoe, can have psychological commitments, by himself, e.g., to trap a goat. But he needs another agent, e.g., Man Friday, to interact with to have S-commitments, e.g., to have dinner at sundown.

2.1 Psychological Commitments

P-commitments are pretty straightforward and are only briefly discussed here. Intentions for future actions are an important concept in AI. Perhaps their salient property is that they involve a commitment on the part of agents. This view has been gaining ground in Philosophy and AI (e.g., see [1, ch. 2], [7, p. 94] and [2, p. 217]). The idea here is that an agent who has an intention is in some way committed to it—not only does he intend to achieve the relevant condition right now, but would also intend to achieve it later, even as the circumstances changed, perhaps for the worse.

Agents can also be committed towards their beliefs, this commitment corresponding to the latter's doxastic or epistemic entrenchment [4, ch. 3]. The more committed an agent is to a belief, the more disposed he will be to not reconsider it, even as he receives new evidence that potentially challenges it or removes its justifications; e.g., if you firmly believe that all birds can fly, you might accept that penguins aren't birds, rather than that some birds cannot fly.

P-commitment entails that the agent continue with a belief that might no longer be justified and an intention that might no longer be feasible or serve the agent's ultimate goals. Thus there is a certain amount of irrationality built into this concept. But for limited agents, commitments can be reconciled with rationality. Limited agents cannot be expected to reason at every moment from "first principles"—this kind of reasoning can be simply too expensive to carry out repeatedly. P-commitments help the agent consider only a few issues, which he can hope to resolve in the limited time he has available. I will just take it that it is a feature of good design of agents that they are able to manage despite these limitations.

2.2 Social Commitments

This kind of commitment arises often in distributed AI, e.g., when the agent promises (to another) to do a certain action. The S-commitment of an agent (to another) in a multiagent system to achieve a certain goal is essential to coordination; e.g., if one agent promises another to be at a rendezvous spot, then the other can take this for granted and be there himself; or if one agent promises to lift one end of a piano, the other can lift the other end, and jointly they can lift the piano, thereby

achieving something they could not have achieved singly. In general, I take S-commitment as the commitment of an agent to another agent, who may be a group that includes the first agent. It is important to note that commitments in classical Distributed Computing are irrevocable—this makes much of this research inapplicable to AI, where such rigidity might be undesirable.

Whether or not S-commitments may be reduced to P-commitments in principle, they are at least in practice an important concept for AI. It would be impossible to design or understand sufficiently complex systems of autonomous agents without invoking the concept of S-commitment. The agents' S-commitments would give us a powerful abstraction with which to view their interactions. Thus, just as psychological concepts are needed to understand complex agents, social concepts are needed to understand systems of agents. And S-commitment is a social concept *par excellence*.

1 **S-commitments as interfaces:** S-commitments are a way of specifying interfaces between intelligent agents in multiagent systems. The most interesting interactions among agents involve communications, whether implicit or implied. Communications can be described fruitfully using social commitment. For example, promises bring into effect a social commitment by the speaker to the hearer; directives presuppose a S-commitment by the hearer to do as told; assertives S-commit the speaker to the statement expressed (even if the speaker is insincere); permissions make the speaker S-committed to allowing the relevant condition to hold; prohibitions presuppose a S-commitment by the hearer to preventing the relevant condition from holding. Interaction protocols between different agents they may be defined so that the interacting agents have the relevant commitments. This generalizes the idea of [13] that participating agents have the requisite intentions and know-how. This idea can also help differentiate communication from other interactions such as resource conflicts. Even though an agent might obtain some information from another because of conflicts, these interactions differ from the properly communicative ones in lacking the properties discussed above.

A practical consequence of this is that one can design agents independently of each other and just ensure that their S-commitments would mesh in properly when they are combined.

2 **S-commitments and individual rationality:** An agent can *bargain* with other agents about the adoption of some S-commitments. If he agrees to do something that another agent intends he do, he might be able to get something in return from him. If he just intended to do something without S-committing to it, he would have no grounds for bargaining; e.g., Crusoe can get Friday to catch a grouse, only by S-committing to milk the goat. Another potential benefit to an agent is that S-commitments can make it easier for the group as a whole to achieve its ends, and this might have some positive utility for him.

3 **S-commitments and group rationality:** The role of S-commitments in group rationality is akin to that of P-commitments in individual rationality. Even when they are not in the immediate (i.e., local) interest of an agent, S-commitments might be a good idea from the systems point of view. When agents in a system make and keep S-commitments, the system can perform better than it might if they each worked selfishly. This too depends on specific systems, but there are numerous real-life systems where S-commitments alone make it possible for any kind of success; e.g., the famous prisoners' dilemma paradox arises when the individually preferred actions of two agents (namely to tell on each other) leads to a worse pay-off for each than the individually less preferred actions of protecting each other. The latter can be done only if the agents are committed to each other to not aid the police.

4 **S-commitments and representation:** S-commitments may be implicit; i.e., they need not be explicitly symbolically represented by the agents, but could instead be derived from their social architectures. Of course, they *could* be symbolically represented by *some* smart introspective agents. Often the relevant agents would be treated only indexically, i.e., relative to the agent's own situation.

E.g., at a stop sign at a street crossing, the other agent would be "the driver to my right," rather than Bill, the councilman. The S-commitment itself to let him go (and the S-commitment on his part to get out of my way) are both a matter of training, here, of learning how to drive. S-commitments may also be generated by the *social roles* of agents; e.g., someone, merely by being a policeman in uniform, is S-committed to chase after a criminal. His not doing so would be a dereliction of duty.

5 **Commitments and coordination:** In traditional construals of group action, the agents in the group are required to have mutual beliefs about the action of the group [8, 5]. Here, it proposed that no such mutual beliefs are necessary. When two agents make social commitments to each other, they are already in a position to perform their joint action. Mutual beliefs are difficult to achieve in realistic scenarios and are highly unstable even when achieved—the slightest suspicion by one of the participants that the other no longer has the relevant belief (to any arbitrary nesting) is grounds for the failure of the mutual belief; e.g., if one agent comes to doubt (for a large natural number, n) that the other believes that he believes ... (n times) the given proposition, their mutual belief no longer exists, even if this doubt were unfounded. This is a fundamental problem for theories that rely on mutual beliefs to explain or define coordination. No number of epicycles added to such a theory can repair this problem—as long as there are mutual beliefs, they will be unstable. Indeed, the requirement of mutual belief may be taken as a purported proof that joint action is impossible (e.g., as in [6]).

A far simpler definition for coordination can be motivated: the involved agents merely have to have the appropriate S-commitments to one another. Two agents are jointly committed to achieving a condition, p, iff each of them is S-committed to the other to achieve a (possibly different) condition such that the achievement of those two conditions would entail p. The discussion below on revoking commitments applies here if an agent needs to drop out of some joint commitment.

2.2.1 Revoking Social Commitments

While P-commitments can be revoked at will (at least in principle), S-commitments cannot be revoked till the agent clarifies this to the concerned agent. Ideally, an S-commitment would be terminated when the requirements it imposed were met (i.e., the commitee is satisfied), or that the commitee released the commiter. These roles make the commitee important.

An agent might also need to give up his S-commitment if the circumstances change, e.g., if his ultimate goals change. As a matter of descriptive fact for humans, the agent might then simply drop his S-commitment and quit acting for it. This kind of unreliable behavior can lead to a number of complications, e.g., as when someone does not show up for an appointment. Prescriptively speaking, we would like to prevent such behavior, at least to a first approximation (we might not always want to prevent it because it might sometimes be acceptable—e.g., someone who foolishly agrees to rob a bank, might be forgiven for not showing up). But in more mundane cases it helps a system if its members do not drop their S-commitments arbitrarily.

We should require that to revoke a S-commitment, the commiter must first get an acknowledgement from the commitee. All that we need is that when a commitment is revoked, the commiter knows that the commitee knows that. Mutual beliefs are not needed since once the commitee believes the commitment is being given up, he need no longer act as if it held. It would be unwise for the commitee to act as if the commitment persisted even though he knows it does not (merely on the hope that the commiter might persist with it a little longer yet). Not requiring mutual beliefs to be established makes for a considerable simplification, as already discussed.

In the full paper, further details are supplied and a formalization of both senses of commitment is proposed that captures their important properties, including their desired interrelationships.

References

[1] Michael E. Bratman. *Intention, Plans and Practical Reason.* Harvard University Press, Cambridge, MA, 1987.

[2] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[3] Daniel C. Dennett. *The Intentional Stance.* MIT Press, Cambridge, MA, 1987.

[4] Peter Gärdenfors. *Knowledge in Flux.* MIT Press, Cambridge, MA, 1988.

[5] Barbara Grosz and Candace Sidner. Plans for discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *SDF Benchmark Series: Intentions in Communication.* MIT Press, Cambridge, MA, 1988.

[6] Joseph Y. Halpern and Yoram O. Moses. Knowledge and common knowledge in a distributed environment (revised version). Technical Report RJ 4421, IBM, August 1987.

[7] Gilbert Harman. *Change in View.* MIT Press, Cambridge, MA, 1986.

[8] H. J. Levesque, P. R. Cohen, and J. T. Nunes. On acting together. In *AAAI-90*, 1990.

[9] John McCarthy. Ascribing mental qualities to machines. In Martin Ringle, editor, *Philosophical Perspectives in Artificial Intelligence.* Harvester Press, 1979. Page nos. from a revised version, issued as a report in 1987.

[10] Munindar P. Singh. Group intentions. In *10th Workshop on Distributed Artificial Intelligence*, October 1990.

[11] Munindar P. Singh. Group ability and structure. In Y. Demazeau and J.-P. Müller, editors, *Decentralized Artificial Intelligence, Volume 2.* Elsevier Science Publishers B.V. / North-Holland, Amsterdam, Holland, 1991.

[12] Munindar P. Singh. Intentions, commitments and rationality. In *Meeting of the Cognitive Science Society*, August 1991.

[13] Munindar P. Singh. Towards a formal theory of communication for multiagent systems. In *International Joint Conference on Artificial Intelligence*, August 1991.