Artificial Intelligence and Law **7:** 97–113, 1999.

An ontology for commitments in multiagent systems:

Toward a unification of normative concepts

MUNINDAR P. SINGH
Department of Computer Science, North Carolina State University, Raleigh, NC 27695-7534, USA
E-mail: singh@ncsu.edu

Abstract. Social commitments have long been recognized as an important concept for multiagent systems. We propose a rich formulation of social commitments that motivates an architecture for multiagent systems, which we dub *spheres of commitment*. We identify the key operations on commitments and multiagent systems. We distinguish between explicit and implicit commitments. Multiagent systems, viewed as spheres of commitment (SoComs), provide the context for the different operations on commitments. Armed with the above ideas, we can capture normative concepts such as obligations, taboos, conventions, and pledges as different kinds of commitments. In this manner, we synthesize ideas from multiagent systems, particularly the idea of social context, with ideas from ethics and legal reasoning, specifically that of directed obligations in the Hohfeldian tradition.

Key words: commitments, multiagent systems, norms.

1. Introduction

Social commitments are crucial to the science of multiagent systems. They are also crucial to the engineering of robust and flexible multiagent systems, e.g., in applications in open systems (Singh, 1997). Although good progress has been made, current theories of commitment fail to cover the full range of normative and organizational phenomena of interest. At the same time, while traditional computer science approaches—as exemplified by distributed databases—have some useful insights, they define commitments as procedurally hard-wired, irrevocable, and therefore quite limiting.

We believe that social commitments in general, and our approach in particular, provide a fruitful point of contact between the multiagent and the legal representation and reasoning—more generally, the ethics and deontic logic—communities. The ethics and deontic logic community can contribute insights gleaned from the long tradition in jurisprudence—as our agents become smarter, legal issues similar to human societies can arise in multiagent settings. Conversely, multiagent metaphors can provide a natural formulation of some longstanding issues in legal reasoning. von Wright argues that the study of deontic concepts should be carried

out as part of *praxeology*, the study of the notions of agency and activity (1968, pp. 12–13). What we propose here may be thought as social praxeology.

Ross motivates the notion of prima facie duties as an alternative to a purely utilitarian view of ethics (1930, pp. 16–22). He develops a range of reasons for such duties, including previous acts such as one's promises or failures resulting in reparation to others, beneficence and nonmaleficence, self-improvement, and justice. Our approach captures similar intuitions through the notion of commitments and metacommitments. We propose a framework called *spheres of commitment*, which incorporates intuitions from information systems to marry commitments with the organizational structure of heterogeneous multiagent systems. Our approach motivates a rich "descriptive ontology" of commitments—to use Castelfranchi's term (1995)—that emphasizes the interplay between commitments and social structure. It defines operations on commitments and groups, distinguishes implicit and explicit commitments, and models social policies as higher-order commitments. To help appreciate our approach, we first lay out our main conceptual assumptions:

A1. Agents can be structured, and are recursively composed of heterogeneous individuals or groups of agents (Singh, 1991a).

A2. Agents are autonomous, but constrained by commitments—or we would have chaos.

A3. Social commitments cannot be reduced to internal commitments, which apply within an agent—the relationships among these concepts cannot be definitional (Singh, 1991b).

A4. Commitments are, in general, revocable; the clauses for revoking them are as important as the conditions for satisfying them.

A5. Commitments arise, exist, are satisfied, revoked, or otherwise manipulated, all in a social context.

A6. Commitments not only rely on the social structure of the groups in which they exist, but also help create that structure.

A7. The semantics of commitments must be distinguished from the pragmatics; what commitments *are* is very different from how they are *used* by agents—it is methodologically critical that the semantics does not replace the pragmatics, or vice versa.

A8. The subjectivist bias of traditional AI must be avoided—commitments and associated conditions are evaluated in the world, not in the mind of any agent, unless of course they refer to an agent's mental state (Singh, 1991b; Castelfranchi, 1995).

The above assumptions are crucial to developing a powerful framework for commitments that can handle the normative concepts in general. For example, many agents are "corporate individuals," to use a term due to Hobbes. The agents are autonomous, but not recklessly so. The commitments they enter into are frequently canceled. However, the creation and cancelation of commitments occurs relative to the prevailing social situation. Lastly, although commitments obviously have a lot to do with the agents' minds, one can speak of their satisfaction or violation independent of any agent.

We believe that assumptions A4, A5, A6, and A7 have not received sufficient attention in the multiagent literature; similarly, assumptions A5, A6, and A7 have not been emphasized in previous work in ethics and legal reasoning. Therefore, this paper will concentrate on defending and using these assumptions. The goals of this paper are to lay out some of the foundational aspects of commitments and associated concepts of groups, organizational structure, and roles. These aspects are then shown to capture many of the key properties of related normative concepts. A part of the challenge, of course, is that normative concepts involve a variety of meanings. For example, Edel identifies 13 readings of "obligation" in the literature on ethics (1961, p. 328). This means that our ontology must be expressive enough to accommodate a wide variety of intuitions.

Section 2 introduces the key ideas in our ontology of commitments. Section 3 introduces spheres of commitment using social policies and structure. Section 4 casts some normative concepts into our framework. Section 5 reviews some of the relevant literature. Section 6 concludes with a discussion of some outstanding issues.

2. Commitments

In order to carry out our program, we need a concept that can serve as a reasonable foundation. We employ *social commitment* as this concept. However, in order to succeed with our program, we need to take a somewhat broader notion of commitment than is taken by other—to our mind, valuable—approaches, such as those of Castelfranchi, von Wright, and Segerberg. However, in our previous work, we have long taken a broader interpretation than the rest of the literature (Singh, 1991b). A pleasant side-effect of taking a broader interpretation is that it helps in unifying several important normative concepts. We will follow our program here, and defer a detailed discussion of the literature till section 5.

Our approach treats commitments as first-class abstract objects with names (Asher, 1993). Naming enables self- and cross-reference among the commitments. We think of commitments as being toward conditions to be achieved rather than actions. In this way, we borrow from the tradition of von Wright (1963), continued in philosophy by Segerberg (1989) and common in much computer science research on actions and intentions, e.g., our own previous work (Singh, 1994). However,

conditions correspond to high-level actions. Using conditions facilitates nesting commitments and constructing "higher-order" commitments (defined below).

An agent's commitments typically constrain him to act in accordance with them. A commitment is *discharged* when the desired condition is obtained. This condition might be a plain physical requirement, or a requirement that some other agent recognizes that some physical requirement is satisfied. To avoid subjectivism, we require that the discharge condition of a commitment be evaluated objectively. However, the condition may explicitly involve the beliefs or other mental states of agents. For example, it is possible to commit to "making the sky green," or to "making the sky appear green to the creditor"—these are different commitments, with different requirements of satisfiability.

Although agents normally act in accordance with their commitments, they sometimes cannot or would not do so. Thus, some commitments must be *canceled*. Indeed, in most settings and applications where agents and multiagent systems are useful, inflexibility is undesirable and the agents must retain some autonomy beyond their commitments. However, commitments should not be canceled arbitrarily, because that would subvert their very purpose. A challenge is to reconcile the apparent tension between these requirements.

2.1. FORM AND CONTENT OF COMMITMENTS

Based on the foregoing, we propose the following abstract representation for commitments. In the remainder of this paper, we use a formal language based loosely on the predicate calculus with special predicates for commitments, etc. We use $\rightarrow$ as a material conditional, and $\Rightarrow$ as a strict conditional. We neglect temporal aspects for simplicity. We do not define the semantics formally, but state constraints on how the various predicates and operators relate to one another.

DEFINITION 1 A commitment is a four-place relation involving a proposition (p) and three agents (x, y, and G). Let $c = \mathsf{C}(x, y, G, p)$ denote a commitment from x toward y in the context of G and for the proposition p. Then, x is the debtor, y the creditor, G the context group, and p the discharge condition of commitment c.

Here, the *debtor* is the agent who is committed, and the *creditor* is the agent who receives the commitment. The creditor need not be the direct beneficiary. For example, in Japanese society, obligations to one's children (beneficiaries) are repayment of debts to one's parents (creditors) (Edel, 1961, p. 330). However, we will assume that the creditor can be treated as the beneficiary for all practical purposes. We revisit this assumption in section 5.

The *discharge condition* is the condition committed to. Each commitment exists in a (social) *context*. Intuitively, the context includes the norms or conventions that apply in the group in which the commitment is instantiated. It provides the court of appeals for adjudicating disputes between the debtor and creditor.

Formally, the context is a group that contains the participating agents, usually in different roles. Groups in general, and context groups in particular, are first-class agents in our framework. Morse mentions in passing how the public may be treated as a plaintiff, i.e., an agent, in law (1995a, p. 227). The idea that the group of agents has sovereignty over its members is basic to democracies. Also, in democracies, the group collectively delegates its powers to its representative, the government—typically, itself a group (Morse, 1995a, p. 221).

Previous approaches to commitments have only looked at context-sensitivity with respect to propositions that define potential exception conditions. For example, you may be obliged to lending someone your car, but not if he is drunk. However, there is also a component of the context that is purely social. This is the component that we emphasize by identification of the context group in our definition.

2.2. OPERATIONS ON COMMITMENTS

Our descriptive ontology includes the following operations on commitments.

O1. *Create* instantiates a commitment; it is typically performed as a consequence of an agent adopting a role or by exercising a social policy (explained below).

O2. *Discharge* satisfies the commitment. We postulate that *discharge* is performed concurrently with the actions that lead to the given condition being satisfied. In this manner, the actions that realize the discharge condition *generate* the *discharge* action (Goldman, 1970).

O3. *Cancel* revokes the commitment, subject to the cancelation clause of the commitment and the policies in effect in the given group.

O4. *Release* essentially eliminates the commitment. This is distinguished from both *discharge* and *cancel*, because *release* does not mean success or failure of the given commitment, although it lets the debtor off the hook. The *release* action may be performed by the context or the creditor of the given commitment.

O5. *Delegate* shifts the role of debtor to another agent within the same context, and can be performed by the new debtor or the context.

O6. *Assign* transfers a commitment to another creditor within the same context, and can be performed by the present creditor (if authorized) or the context.

Interestingly, except for *discharge*, these can be, and typically are, performatives (Austin, 1962) of the debtor, creditor, or context. This is because the debtor can

unilaterally create or cancel commitments by performing the right illocutions (under appropriate circumstances), but cannot always discharge a commitment by just saying so. As a reviewer correctly pointed out, the commitment can be to perform a communicative act, in which case it can be discharged by a performative. But there is a nesting of the actions in such a case.

We use the names of the operations as predicates in our language. Thus, $cancel(x, c)$ denotes a proposition, which is true precisely when agent x cancels commitment c. This presupposes that x is the debtor of c.

3. Spheres of Commitment

We postulate two kinds of agents: named individuals and named groups. Individuals are unstructured; groups are constructed from individuals or other groups by specifying a social structure. By naming groups we can allow their membership to change while maintaining a constant identity. This enables us to talk of the commitments relating to such agents. The same set of agents may form different groups, potentially with different structures. A *sphere of commitment (SoCom)* is a group viewed in conjunction with its *roles* and their concomitant commitments.

3.1. STRUCTURE

We consider the following operations with respect to groups. A group may be *created*; an agent may *adopt* a role; an agent may *reassign* himself to another role; an agent may *exit* a group. Not all of these operations may be enabled or voluntarily performed, of course. For example, in a caste-based society, each agent is assigned a role automatically and cannot change a role or exit the society, except by death. Most current applied multiagent systems are similar! There is, of course, research into societies defined in terms of agents playing different roles, and entering or exiting roles to reorganize societies. For example, Glaser and Morignot (1997) study reinforcement learning for reorganizing societies, although they do not focus on commitments *per se*.

3.1.1. *Explicit versus Implicit*

We previously proposed that the structure of a group is given by the constraints on the interactions among its roles (Singh, 1991a). We elaborate the interactions somewhat differently now. We distinguish two main kinds of commitments—*explicit* and *implicit*—which are intimately related to two main kinds of interactions, termed *strategic* and *reactive* in (Singh, 1991a). Explicit commitments are explicitly represented by one or more of the agents; implicit commitments are not. Consequently, explicit commitments can feature in the agents' communications. Intuitively, implicit commitments are those that the agents do not need to articulate, but which are implicitly common knowledge (or mutually believed) in the system. This is important, because the only kind of common knowledge in a real system is

what is present without having to be created by communications (Chandy & Misra, 1986).

Explicit interactions rely for their meaning upon explicit commitments among the communicating agents. Consider the different classes of illocutionary acts (Austin, 1962; Searle, 1969). For example, commissives ordinarily bring into effect a commitment by the speaker to the hearer. This commitment is as much of a commitment as any other that the speaker may have. Indeed, we strongly agree with Castañeda, when he states that "each act of promising is an act of the same type as an act or process of enactment of a law" (1975, p. 181).

Directives presuppose a commitment by the hearer to do as told; in order to succeed, they lead to a specific commitment by the hearer. Assertives commit the speaker to the statement expressed. Permissives make the speaker committed to allowing the relevant condition to hold or to specifically release a prior commitment of which the speaker is the creditor or context group.

Implicit interactions rely for their meaning upon implicit commitments. The implicit commitments correspond to "habits" of interaction that lead to the observed—arguably, the "correct"—behavior of the system without necessarily any explicit representation or reasoning by the interacting agents. An example is a commitment not to break into a line at a bus stop. There is of course a fine line between implicit and explicit commitments, because implicit ones can become explicit when discussed. But the distinction is conceptually and practically an important one to maintain.

3.1.2. *Flow down*

Each role comes with its commitments. In some cases, the commitments that an agent acquires because of adopting a particular role can be overridden, possibly by commitments acquired through another role. However, in some cases, they cannot—the agent must give up the original role or exit the original group in order to be released from the commitments that came with the original role. We call this situation the *flow down* of commitments. Flow down is related to conflict of interest situations in which an agent has somehow ended up with conflicting roles. For example, an agent may not be able to carry out the commitments of the manager role if one of the staff being managed is his son.

3.2. POLICIES

Social policies are conditional expressions involving commitments and actions on commitments. Social structure and social policies are two faces of the same coin. The former applies within a group; the latter apply across agents, including those who constitute a group.

DEFINITION 2 A formal expression is 0-order iff it does not refer to a commitment, and is $(i + 1)$-order iff the highest order commitment it refers to is i-order.

DEFINITION 3 A commitment $c = \mathsf{C}(x, y, G, p)$ is i-order iff the order of p is i.

Consequently, policies for i-order commitments are $(i + 1)$-order expressions. Policies have a computational significance, which is that they can lead to executing various operations on commitments, even without explicit reference to the context group. It is their locality that makes policies useful in practice.

Policies may apply to each of the operations on commitments that were defined above. In fact, policies on the operations are routine ways of specifying the constraints on a given commitment or class of commitments. To simplify the presentation, we use the term *release policy*, etc. to refer to a social policy that applies to the performance of a *release* (or other appropriate) operation on the given commitment.

In a previous version, we included an explicit *cancelation clause* as part of the definition of commitment. The cancelation clause was the enabling condition under which the commitment could be canceled. Setting it to false made the commitment irrevocable; setting it to true allowed the commitment to be given up at will. Most interesting cases lie in between, e.g., setting the cancelation clause of commitment c to *canceled*(c') means that commitment c' must be canceled in order to cancel c. We retain the same intuitions in the present version. However, for reasons of elegance and to simplify the definition, we now treat cancelation just like any other operation on commitments. Thus, a cancelation clause of the previous version shows up as a cancelation policy now.

3.3. NORMATIVE VERSUS NON-NORMATIVE POLICIES

Social policies are policies on commitments. They may or may not be shared by all agents, and they may or may not be norms of the given society. It is possible to have non-normative policies. For example, an agent, x, might adopt a policy of being altruistic, which might be realized as adopting commitments based on requests. Thus we have $(\forall y, G, p : y \in G \rightarrow \mathsf{C}(x, y, G, q))$, where $q \triangleq (\text{request}(y, x, p) \Rightarrow create(x, \mathsf{C}(x, y, G, p)))$. Now when some agent y requests that x bring about or perform p, x adopts a commitment for p. However, x is not committed to any other agent to be altruistic, and it is not a social norm in this society that x be altruistic.

However, in most interesting cases, social policies are norms. In such cases, they are themselves commitments of the agents. A normative situation related to the above example is when x fills a specific role in group G, where the designated role is supposed to be altruistic toward the other members of G. In that case, x would have a commitment to be altruistic. That is, we have a higher-order commitment $\mathsf{C}(x, G, G, (\forall y, p : y \in G \rightarrow \mathsf{C}(x, y, G, q)))$, where the discharge condition of

this commitment is precisely the policy of altruism given above. Similar policies can be defined where the debtor is the context group, and the discharge condition is performing an action such as *release* of a commitment.

4. Relating Normative Concepts and Commitments

We now discuss the relationships between commitments as we understand them, and traditional concepts that occur in the literature.

4.1. TRADITIONAL CONCEPTS

We now refer back to some pretheoretic normative notions, and show how they can be mapped to the concepts introduced above. These notions are informal, or at least understood in different ways by different authors. We give an example after the enumeration.

N1. *Pledge*. Since commitments in our framework are more powerful than usual, we also identify a *pledge* as an explicit commitment. Pledge, in this sense, is a commitment, not a commissive performative. In other words, $\mathsf{P}(x, y, G, p) \triangleq (\exists c : c = \mathsf{C}(x, y, G, p) \& \text{explicit}(c))$. In general, all six operations are allowed on pledges. Pledges are often, but not always, the result of commissive performatives—they can also arise through conventions or other kinds of higher-level commitments.

N2. *Ought*. We believe that the classical approaches in deontic logic, which express *ought* in absolute terms, are unsuitable, because *ought* is inherently contextual. By contrast, we analyze *ought* as relativized to the context group—thus *ought* corresponds to a commitment whose creditor is identified with its context group, and whose cancelation policy is set to false. That is, $\mathsf{O}(x, G, p) \triangleq (\exists c : x \in G \& c = \mathsf{C}(x, G, G, p))$.

We can, of course, also represent the traditional deontic *ought*, although we do not approve of it. That is, $\mathsf{O}_d(x, p) \triangleq (\exists G : x \in G \& \mathsf{O}(x, G, p))$. $\mathsf{O}_d(x, p)$ flows down.

Interestingly, *ought*s cannot be canceled, delegated, or assigned. Traditionally, cancelation is effected through the use of conditional or dyadic obligations. $\mathsf{O}(x, G, p)$ can be released by its context—in this way, the context encodes the exception conditions. However, $\mathsf{O}_d(x, p)$ cannot be released—it lacks both a context and a creditor. The above definition respects Castelfranchi's syllogism that a commitment implies that the debtor ought to discharge it (1995).

N3. *Taboos* are implicit 0-order commitments relative to a context. That is, $\mathsf{T}(G, p) \triangleq (\forall x \in G : (\exists c : c = C(x, G, G, \neg p)$ & implicit(c) & order$(c, 0)$ & $(cancel(x, c) \Rightarrow \mathsf{false})))$. Taboos apply uniformly to all members of a group. They are implicitly created and must be discharged—no other operation applies, and they cannot be overridden. Not satisfying them can lead to excommunication from the given group. Taboos, being implicit, are not talked about in the group.

N4. A *convention* or *custom* is also relativized to a context and creditor, but is implicit. We require conventions to be higher-order, because they lead to lower-level commitments when exercised; the resulting commitments may be explicit or implicit. That is, $\mathsf{Conv}(G, p) \triangleq (\exists c, \exists i, \forall x \in G : c = \mathsf{C}(x, G, G, p)$ & implicit(c)&order(c, i)&$(cancel(x, c) \Rightarrow \mathsf{false})$&$i > 0)$. We distinguish conventions from taboos in that they typically lead to other lower-order commitments. An example of a convention is always responding to whoever communicates with you.

N5. A *collective commitment* of a group of agents may be defined as the conjunction of the commitments of the individuals to the group, in the context of the same group, and which can be given up only if the members mutually believe that it is impossible. Dunin-Keplicz & Verbrugge study this and related notions, also summarizing some previous work (1996). Mutual belief in a group that q holds means roughly that each member of the group believes q, and that the others believe q, and so on, to arbitrary levels of nesting (Fagin *et al.*, 1995). (Barwise (1989a) compares three alternative formulations of mutual beliefs, but these are essentially similar for our purposes.) Here we use MB to notate mutual belief. Neglecting temporal aspects, we have $\mathsf{Coll}(G, p) \triangleq (\forall x \in G : \mathsf{C}(x, G, G, p)$&$(cancel(x, c) \Rightarrow \mathsf{MB}(G, \neg p)))$.

N6. In our understanding, *obligation* has two main readings: (a) one close to *pledge*, and (b) the other close to *ought*. Therefore, we shall not discuss it separately.

In the above manner, our approach can model several normative concepts in a unified framework. These concepts may involve mental states as in N5 above, or may not. The flexibility of our approach can enable a declarative specification of the normative behavior of multiagent systems in a variety of applications.

To use an example suggested by a reviewer, suppose Andy accidentally damages my car in a parking lot. If I had lent him the car, there might be an explicit pledge from him to repair any damage. If not, he might follow the convention in our society that he should repair the car. Lastly, Andy might satisfy the personal requirement

that "good people ought to fix whatever they damage," which can lead him to adopt the lower-order commitment to perform the repairs. This is Andy's policy, but may not be a norm in our, let's say aggressive, society.

4.2. HOHFELDIAN CONCEPTS

The concepts in N7, N8, N9, and N10 are due to Hohfeld (1919), and carry a clearer meaning in the literature. In the following, G is a group corresponding to the entire society or a legally relevant subset thereof, such as a city or a county.

N7. A *claim* or *right* is what an agent can demand from another. It is like a commitment with respect to the relevant context, which is not made explicit by Hohfeld. Thus, we have $\mathsf{Claim}(x, y, p) \triangleq \mathsf{C}(y, x, G, p)$.

N8. A *privilege* is a freedom an agent has from claims of another. In other words, it is an absence of a duty to refrain from the given act. In this sense, a privilege is the dual of a claim with the roles of the agents reversed, i.e., $\mathsf{Priv}(x, y, p) \triangleq \neg\mathsf{Claim}(y, x, \neg p)$.

N9. A *power* refers to the ability of an agent to force (if he so desires) the alteration of a legal relation in which the other agent participates. Thus, we have $\mathsf{Power}(x, y, r) \triangleq \mathsf{C}(G, x, G, \mathit{request}(x, G, r) \Rightarrow \mathit{perform}(G, r))$, where r is an operation on commitments whose creditor or debtor is y, and context group is G. In other words, x has the power to alter y's legal relations by requesting the context to perform an operation that changes the relevant legal relations of y. Examples of such changes include taking away property belonging to y or granting him property he didn't previously have.

N10. An *immunity* means a freedom from the power of another agent. Thus, $\mathsf{Immunity}(x, y, p) \triangleq \neg\mathsf{Power}(x, y, \neg p)$.

The concepts in N7, N8, N9, and N10 are the core Hohfeldian concepts. Our terminology follows Kanger in using *claim* for *right*, and *exposure* for *no-right* (1971). The *correlate* of a concept is obtained by switching the roles of x and y in its definition. Duty, exposure, liability, and disability are the correlates of claim, privilege, power, and immunity, respectively. The above definitions support the observation that claim and privilege, and power and immunity are duals of each other with the roles of the agents reversed.

Our definition of power bases an agent's ability to change legal relations on the corresponding and primary such power vested with the context group. This reflects our philosophical assumption of the ultimate power of the context group. By showing power and immunity as affirming or negating metacommitments, we also reflect the legal intuition that claim and privilege are a more direct component

of Hohfeld's analysis than power and immunity, a point made by Fuller, cited in (Morse, 1995a, pp. 244–245).

Hohfeld argued that "strictly fundamental legal relations" cannot be satisfactorily formalized (p. 36). We showed how these fundamental relations can be replaced by the single one of commitment, which although not formally defined can help unify the definitions of the others.

4.3. NORMS AND COMMITMENTS

So, when all is said and done, how do norms relate to commitments? In our view, commitments—of order greater than 0, i.e., metacommitments—create a society. The metacommitments are the norms of this society. In the context provided by this society, lower-level commitments are instantiated due the actions of the different agents, leading to cohesive interactions among them. Some of the lower-level commitments—if they are metacommitments—would themselves be norms. However, these can also be social policies (adopted by different agents) that are not norms. For example, the altruistic agent of section 3.3 acquires commitments based on his personal policy. In this way, commitments can arise as duties of justice and kindness, or even self-improvement described by Ross. Consequently, we believe we have supported the thesis that commitments are both the progenitors and the descendants of norms. However, they are not exclusively the descendants of norms.

5. Literature

As should be clear by now, our work seeks to synthesize ideas from multiagent systems on the one hand, and ethics and legal reasoning on the other. We made some allusions to the literature in the foregoing, but we discuss some important works in more detail below.

5.1. MULTIAGENT SYSTEMS

Although much research has been performed on commitments and social relations, for reasons of space, we shall compare our approach primarily to Castelfranchi (1995), who also considers multiagent system issues. We share many intuitions with Castelfranchi, but emphasize them to different degrees. In particular, we agree on the limited autonomy of agents, their dependence relations with each other, the irreducibility (in general) of social concepts to mental concepts, and the normative nature of roles. However, Castelfranchi requires stronger relationships among these concepts, which we believe can hold only in specific cases. For example, he requires that the creditor have the committed condition as a goal, and further that this is mutually known to the creditor and debtor. In addition, the creditor should commit to accepting the commitment.

Besides potential problems about how these recursive commitments might bottom out, this appears too strong in several cases. For example, if I offer a student a research assistantship, I am committed as soon as I mail the letter—she might decline it or come back later (within some designated interval) to claim it. In this case, the commitment holds even though the mutual belief does not arise much later (if ever). If the letter is lost, I can get off the hook, but only because the context group and creditor aren't aware of my commitment. Indeed, by exhibiting the letter, anyone can prove that the commitment held.

Castelfranchi also requires that the debtor be committed to not opposing the creditor's complaints if the commitment is canceled; in our case, this is not required by the commitment itself, but may or may not be a social policy. In some settings, the debtor in default may be expected to accept any complaints meekly, but in others he may not. Castelfranchi's notion of implicit commitments is different from ours in that he defines them as essentially defeasible commitments—if there is an expectation and the (potential) debtor does not deny it, then he becomes committed. This is certainly a useful notion, although we would capture it through social policies governing the creation of commitments through, on the one hand, accepting requests, and on the other hand, refraining to reject them. For us, the implicit commitments are acquired by the agents based on their membership in a group. Such commitments are usually not up for debate.

Castelfranchi's notion of *ought* resembles traditional deontic logic in that it is not relativized to any context. Further, he does not exploit the structure inherent in agents and organizations to the extent that we do. He does not assign as much discussion to the cancelation of commitments.

Sichman *et al.* develop a theory and interpreter for agents who can perform social reasoning (1994). Their agents represent knowledge about one another to determine their relative autonomy or dependence for various goals. Dependence leads to joint plans for achieving the intended goals. This theory does not talk about commitments *per se*, so it is complementary to our approach. Levesque *et al.* reduce commitments to mutual beliefs, thus suffering from the problems indicated above (1990). They also hardwire a specific approach to canceling commitments (for joint intentions)—the participating agents must achieve a mutual belief that the given commitment is off. Jennings postulates *conventions* as ways in which to reason about commitments (1993). For teams, he requires a "minimum" convention, which recalls Levesque *et al.*'s approach.

5.2. DEONTIC LOGICS

Traditional approaches tend to focus on single-agent issues. For example, von Wright considers a single agent "alone with nature" with at best indirect reference to other agents and interactions (1968, pp. 48–49). This we believe is fundamentally limiting because of the multiagent nature of the phenomenon being studied.

Traditionally, largely as a result of their single-agent focus, deontic logics have studied obligations that are not directed, and have considered commitments as conditional obligations of some form. For example, von Wright defines a commitment as an obligation toward a conditional proposition of the form $O(p \rightarrow q)$ (1968, p. 77) with the view that the performance of p commits the agent to the performance of q. Prior defines commitment as $p \rightarrow Oq$ (cited in (Føllesdal & Hilpinen, 1971, p. 24)). Segerberg considers commitments as expressed in locutions such as "by virtue of A, g commits h to B," where g and h are agents and A and B are propositions (1971, p. 148). However, the commitments in this case are not directed and the agents are suppressed in all further technical development. As in the abovementioned approaches, the commitment is really a conditional obligation. Segerberg's definition of commitments as $\Box(p \rightarrow Oq)$ formalizes this idea in a strict version of Prior's idea (p. 154).

Although much of the work on obligations, commitments, and related normative concepts has been carried out in an implicitly single-agent framework, there are some notable exceptions. Specifically, in the work descending from the Hohfeldian tradition, the notion of direction is explicit. The original reference to that tradition is (Hohfeld, 1919); Morse also gives a summary in his tribute to Hohfeld (1995a). Kanger was among the first to formalize Hohfeld's ideas (1971). However, his development didn't do justice to the directedness of the original concepts. This is more obvious from (Kanger, 1985), and has been pointed out by several others, including most recently Herrestad & Krogh (1995).

We have some additional objections to Kanger's approach (1971). He defines *ought* and its dual *right*, roughly analogous to obligatory and permissible, and uses these to formalize Hohfeld's primitives. Duty means that the agent ought to; privilege means that the agent would be right not to do the negation; power means that the agent has a right to; and disability means that the agent ought not to do the negation (pp. 42–44). The first two are acceptable. However, the formalizations of power and disability do not do justice to Hohfeld's original meanings. In Hohfeld's terms, power and disability are about altering legal relations. Hohfeld defines *operative facts* as those that, under legal rules, suffice to change legal relations (1919, p. 32). Thus Kanger is susceptible to the charge of confusing operative facts with alterations in legal relations proper, which Hohfeld anticipated and warned against (p. 52).

Alchourrón and Bulygin consider permissions as either lifting prohibitions or issuing permissive norms (1981, pp. 116–117). We believe that these notions are closely related to the Hohfeldian concepts. Lifting a prohibition appears to correspond to removing a duty, which is identical to granting a privilege. Issuing a permissive norm appears to correspond to granting an immunity, which is identical to removing a liability.

Herrestad & Krogh propose a formalization of directed obligations, highlighting some interesting issues in legal representation. Herrestad & Krogh reduce a directed obligation from i to j to do A to the conjunction of a bearer-relativized

obligation on *i* to do *A*, and a counter-party-relativized preference for *j* that *i* do *A* (1995, §4). In other words, the debtor is obliged to do the given action, and the creditor is obliged to prefer its occurrence. This recalls the idea of Castelfranchi that the creditor accept the commitment. We agree that in many common uses of commitments, the creditor will indeed favor receiving the given commitment. However, we take this as a question of pragmatics, not of semantics. Indeed, if the creditor does not prefer the commitment, he can easily release the given commitment (provided the applicable release policy allows). Thus a debtor—at least in cooperative settings—would not carry out a commitment that the creditor did not desire.

It is technically simpler—often, more elegant—to use a simpler definition in the semantics. For example, Herrestad & Krogh observe that their stronger definition of directed obligation leads to a weaker definition of permission. When they repair the definition, they again insert in it the requirement that a permission holds only if the agent against whom the permission is directed prefers for the permitted action not to be performed. This too is a question of pragmatics for surely permissions that no one is opposed to might not be worth talking about.

To appreciate our distinction between pragmatics and semantics consider the requirement that only those propositions can be true that someone believes to be false. Or, alternatively, an agent believes something true only if another agent believes it false. As (semantic) requirements on truth or belief, these are silly. Yet, as (pragmatic) requirements on facts and beliefs that are worth talking about by the agents or by their designers, these are eminently reasonable.

6. Conclusion and Future Work

As von Wright observes, an agent may err and deviate from the "path of righteousness" after which it might recover through additional actions (1968, p. 74). Traditional deontic logics don't permit such actions, leading to the *contrary-to-duty* paradox, and others of its ilk. We made these deviations and corresponding recovery actions a central theme for our approach. We did not explore the connection between social commitments and rationality. This is especially important in cases of conflict, where the *strength* of a commitment may need to be modeled. Some of these issues are being studied by Boman (1996).

Our approach highlights the interplay between commitments and the structure of multiagent systems. The agents always commit in the context of their multiagent system, and sometimes to that system. The multiagent system serves as the default repository for the normative social policies, although these can be assigned to the member agents. In this manner, we can reconcile the tension between flexibility and commitment, and between autonomy and coherence. We recognize that there is need for a model-theoretic semantics to underpin the proposed notion of commitments. At the present, we have been concerned mainly with conceptual issues, and we leave the development of a semantics to future research.

Our approach does not attempt to reduce social commitments to mental concepts such as mutual beliefs, which require the agents to hold beliefs about each other to unbounded levels of nesting. It is well-known, and we have long argued that, mutual beliefs cannot be implemented except with strong additional simplifying assumptions, which is why the direct approach of using social constructs is more appealing (Singh, 1991b). We conjecture that named groups and commitments can provide the necessary connections among the agents.

We also conjecture that as groups become established, their interactions become routinized and their commitments become implicit. This leads to greater efficiency and reduced flexibility. Lastly, we conjecture that commitments flow down more easily when they lack a context, whereas the presence of a context gives an opportunity to flexibly override the stated requirements.

Acknowledgements

This work is supported by the NCSU College of Engineering, the National Science Foundation under grants IIS-9529179 and IIS-9624425, and IBM corporation. I am also deeply indebted to the anonymous reviewers for their helpful comments, which have improved this paper considerably.

References

Alchourrón, Carlos E. & Bulygin, Eugenio (1981). The expressive conception of norms. In: *(*Hilpinen, 1981*)*. 95–124.

Asher, Nicholas M. (1993). *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht, Holland.

Austin, John L.; (1962). *How to Do Things with Words*. Clarendon Press, Oxford.

Barwise, Jon (1989a). On the model theory of common knowledge. In: *(*Barwise, 1989b*)*. 201–220.

Barwise, Jon (1989b). *The Situation in Logic*. Center for the Study of Language and Information, Stanford.

Boman, Magnus (1996). Implementing norms through normative advice. In: *Proceedings of the International Conference on Multiagent Systems (ICMAS) Workshop on Norms, Obligations, and Conventions*.

Castañeda, Hector-Neri (1975). *Thinking and Doing: The Philosophical Foundations of Institutions*. D. Reidel, Dordrecht, Holland.

Castelfranchi, Cristiano (1995). Commitments: From individual intentions to groups and organizations. In: *Proceedings of the International Conference on Multiagent Systems*. 41–48.

Chandy, K.M. & Misra, Jayadev (1986). How processes learn. *Distributed Computing* **1**, 40–52.

Demazeau, Yves & Müller, Jean-Pierre (eds) (1991). *Decentralized Artificial Intelligence, Volume 2*. Elsevier/North-Holland, Amsterdam.

Dunin-Keplicz, Barbara & Verbrugge, Rineke (1996). Collective commitments. In: *Proceedings of the International Conference on Multiagent Systems*. 56–63.

Edel, Abraham (1961). Science and the structure of ethics. In: *(*Neurath *et al.*, 1970*)*. University of Chicago Press. 273–377.

Fagin, Ronald, Halpern, Joseph Y., Moses, Yoram & Vardi, Moshe Y. (1995). *Reasoning About Knowledge*. MIT Press, Cambridge, MA.

Føllesdal, Dagfinn & Hilpinen, Risto (1971). New foundations for ethical theory. In: *(*Hilpinen, 1971*)*. 1–35.

Glaser, Norbert & Morignot, Philippe (1997). The reorganization of societies of autonomous agents. In: *Proceedings of the 8th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW)*. 98–111.

Goldman, Alvin I. (1970). *A Theory of Human Action*. Prentice-Hall, Englewood Cliffs, NJ.

Herrestad, Henning & Krogh, Christen (1995). Obligations directed from bearers to counterparties. In: *Proceedings of the 5th International Conference on Artificial Intelligence and Law*. 210–218.

Hilpinen, Risto (ed.) (1971). *Deontic Logic: Introductory and Systematic Readings*, volume 33 of *Synthese Library*. D. Reidel, Dordrecht, Holland.

Hilpinen, Risto (ed.) (1981). *New Studies in Deontic Logic: Norms, Actions, and the Foundations of Ethics*, volume 152 of *Synthese Library*. D. Reidel, Dordrecht, Holland.

Hohfeld, Wesley Newcomb (1919). *Fundamental Legal Conceptions as Applied in Judicial Reasoning and other Legal Essays*. Yale University Press, New Haven, CT. A 1919 printing of articles from 1913.

Holmström, Ghita & Jones, Andrew J.I. (eds.) (1985). *Action, Logic and Social Theory*. Societas Philosophica Fennica, Helsinki.

Huhns, Michael N. & Singh, Munindar P. (eds.) (1998). *Readings in Agents*, Morgan Kaufmann, San Francisco.

Jennings, N.R. (1993). Commitments and conventions: The foundation of coordination in multi-agent systems. *Knowledge Engineering Review* **2**(3), 223–250.

Kanger, Stig (1971). New foundations for ethical theory. In: *(*Hilpinen, 1971*)*. 36–58.

Kanger, Stig (1985). On realization of human rights. In: *(*Holmström & Jones, 1985*)*. 71–78.

Levesque, H.J., Cohen, P.R. & Nunes, J.T. (1990). On acting together. In: *Proceedings of the National Conference on Artificial Intelligence*. 94–99.

Morse, H. Newcomb (1995a). Hohfeld. In: *(*Morse, 1995b*)*. Chapter 10, 213–245.

Morse, H. Newcomb (1995b). *The Thinkers*. University Press of America, Lanham, MD.

Neurath, Otto, Carnap, Rudolf, & Morris, Charles (eds.) (1970). *Foundations of the Unity of Science: Toward an International Encyclopedia of Unified Science*, volume II. University of Chicago Press, Chicago and London. Originally published as separate monographs.

Ross, William D. (1930). *The Right and the Good*. Oxford University Press, Oxford.

Searle, John R. (1969). *Speech Acts*. Cambridge University Press, Cambridge, UK.

Segerberg, Krister (1971). Some logics of commitment and obligation. In: *(*Hilpinen, 1971*)*. 148–158.

Segerberg, Krister (1989). Bringing it about. *Journal of Philosophical Logic* **18**, 327–347.

Sichman, Jaime Simão, Conte, Rosaria, Demazeau, Yves, & Castelfranchi, Cristiano (1994). A social reasoning mechanism based on dependence networks. In: *Proceedings of the 11th European Conference on Artificial Intelligence*. 188–192. Reprinted in Huhns & Singh (1998).

Singh, Munindar P. (1991a). Group ability and structure. In: *(*Demazeau & Müller, 1991*)*. 127–145.

Singh, Munindar P. (1991b). Social and psychological commitments in multiagent systems. In: *AAAI Fall Symposium on Knowledge and Action at Social and Organizational Levels*. 104–106.

Singh, Munindar P. (1994). *Multiagent Systems: A Theoretical Framework for Intentions, Know-How, and Communications*. Springer-Verlag, Heidelberg.

Singh, Munindar P. (1997). Commitments among autonomous agents in information-rich environments. In: *Proceedings of the 8th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW)*. 141–155.

von Wright, Georg Henrik (1963). *Norm and Action*. Routledge & Kegan Paul, London.

von Wright, Georg Henrik (1968). *An Essay in Deontic Logic and the General Theory of Action*, volume 21 of *Acta Philosophica Fennica*. North-Holland, Amsterdam.