Chapter 31
Commitments in multiagent systems
Some history, some confusions, some controversies, some prospects

Munindar P. Singh

Abstract The notion of commitments as a foundation for understanding interactions among agents has been under development for about twenty years. Cristiano Castelfranchi has contributed to clarifying the conception of commitments by bringing in insights from social psychology. In this essay, I briefly review the conceptual development of commitments in multiagent systems, identifying the key themes and some lingering confusions. I also highlight some ongoing debates with Castelfranchi and some promising directions for future research.

1 Introduction

Cristiano Castelfranchi writes about agents like Michelangelo painted his frescoes. No, I don't mean to suggest that Cristiano writes lying precariously on his back on scaffolding twenty meters above the floor–although one can never be too sure about the ways of Italian intellectuals. Seriously, though, I do mean to suggest that Cristiano naturally envisions and describes complex scenes with many characters and details. The effect is beautiful indeed.

My goal in this short essay, by contrast, is to come to these frescoes as a computer scientist–generally, focusing on a few characters and their particular details in an attempt to understand some components of the scene better.

Professor Castelfranchi has made varied and numerous contributions to identifying, developing, and popularizing the social perspective on multiagent systems. Specifically, I want to focus on the notion of commitments, which Professor Castelfranchi and I have been contemplating and discussing for nearly two decades [3, 4, 38]. This is not to suggest that others haven't contributed to this topic–the study of commitments has become a veritable

Munindar P. Singh
Department of Computer Science, North Carolina State University, USA
e-mail: singh@ncsu.edu

cottage industry–but merely to focus the presentation on themes that interest Professor Castelfranchi and me the most.

2 A Brief Retrospective

Commitments in multiagent systems turn out to be quite different from "commitments" as have long been discussed in artificial intelligence (AI) and philosophy. In traditional AI work, a commitment was understood as the commitment of a single agent to some belief or to some course of action. For example, the AI planning literature of the 1970s advocated an approach called least-commitment planning [35] wherein a planner (working as part of or on behalf of a single agent) would create a plan that left as many of the options of the agent open as it could and for as long as it could. That's a fine idea for a single-agent setting. Notice–as an aside–that in such a setting a commitment is not quite desirable–an agent is best off when its commitments are minimized.

In the mid to late 1980s, when research began in earnest on multiagent systems, researchers adopted the notion of commitment as a way to understand organizations of agents. A commitment in a multiagent system captures a relationship between two parties. A traditional planning-style commitment to one's plans would not suffice. Even though the researchers recognized this, partly because they came from an AI background, they came to the notion of commitments with an attendant mentalist bias [30]. Thus they distinguished multiagent commitments from planning commitments, but only to the extent of somehow reducing multiagent commitments to combinations of mutual beliefs and intentions. A mutual belief between two or more agents is a proposition p where each believes p and each believes that each believes p, and so on, to arbitrary nesting [8]. Mutual or joint intentions are similar in spirit, though somewhat more subtle [30]. In other words, traditional researchers retained their mentalist perspective but hoped that the mutuality of the beliefs or intentions would provide the glue between the agents.

But, as Professor Castelfranchi has eloquently and forcefully argued, social relationships are irreducible to the mental attitudes. And especially in multiagent systems we are concerned with the modeling and enactment of interactions of autonomous and heterogeneous agents. Thus commitments can easily exist or fail to exist with or without any beliefs or intentions on part of any of the agents. I return to this topic in Section 4 along with addressing some other confusions.

In contrast, the social and organizational metaphors provided a more straightforward way to think of multiagent systems, and especially as a way to formulate commitments. It has been long known that human organizations develop and apply standard operating procedures–as, for example, explained by Herbert Simon [37]. And, especially in settings where there may be no mathematical guarantee of obtaining a rigorously correct state or outcome, applying a standard operating procedure would be the rational way for an organization to proceed–in essence, we would define the state or outcome

emerging from such a procedure as being correct. A member of the organization, when faced with a particular situation, could act according to any of the applicable standard operating procedures. Even if the particular outcome in that situation turned out to be undesirable or even harmful, the member would generally not be considered as having been in violation. For example, if a patient collapses in an apparent heart failure, a paramedic nurse may be expected to give the patient a shot of nitrates. The nurse would be deemed to have done the right thing if he gives the patient the specified amount of the recommended medication even if he is unable to save the patient's life or saves the patient's life but inadvertently causes other complications. Clearly there are cases where the standard expectations may be higher and a member of an organization would need to select an appropriate operating procedure in order to avoid all blame. Further, the expectations can vary depending upon the role and qualifications of the member involved. In the above example, we may expect an emergency physician or a cardiologist to consider additional information and potential risks beyond what we expect a nurse to consider in deciding a course of action. But regardless of whether we consider a simplified notion of an operating procedure or a more complex one, the common feature is that the organization empowers its members to act in circumstances that are far from ideal.

To me, the foregoing line of thought led to an inkling of an idea that an organization be able to commit to a course of action. More pertinently, the commitment here arose from the member to the organization. Such thinking led me to distinguish two kinds of commitments: (1) an internal one, which I then termed psychological or P-commitment and (2) an external one, which I termed social or S-commitment [38]. Psychological commitment is the standard concept in AI. Social commitment is the concept that we now refer to as commitment in the multiagent systems community. The AI researchers resisted social commitments. I am grateful to Professor Castelfranchi for lending his support to this area when it was emerging.

Social commitments have some interesting features distinguishing them from psychological commitments. First, a social commitment is directed from one party (its *debtor*) to another (its *creditor*). This terminology reflects the intuition that the debtor is committed to doing something for the creditor.

The idea was to distinguish this from the more obvious notion of a beneficiary. Specifically, a commitment may be directed toward one party but the beneficiary might be another. For example, a shipper may commit to a merchant to deliver a package to a customer. Here the shipper would be the debtor and the merchant the creditor. The apparent beneficiary, the recipient of the package, may show up only within the body of the condition that the shipper commits to bring about. Notice that the logical form of the above commitment is the same as of the commitment where the local police constable commits to the district attorney to deliver a subpoena to or arrest a citizen. We would generally not think of the citizen being subject to a subpoena as being a beneficiary. For this reason, it is advisable to leave the value judgments of who is the beneficiary and who is not outside of the general concept of commitments. Indeed, such value judgments are often accompanied by presumptions about

various psychological concepts, which too we ought to minimize in the general theory.

The second interesting, and even less common, aspect of social commitments was the idea of incorporating an *organizational context* into the notion of commitments. The organizational context of a commitment describes the organization or "system" in which the commitment arises, providing support of the normative backdrop for commitments and interactions among autonomous parties. The debtor and creditor of the commitment would thus generally be members of the context organization. An example would be a commitment from a seller to a buyer operating within the eBay marketplace wherein the seller is committed to shipping some goods. That commitment references eBay as its organizational context. Here, eBay might penalize a seller who doesn't discharge the commitment.

A related intuition is that agents can be composed. In other words, what, from one perspective, appears to be an individual entity and functions as one entity (and therefore is a well-defined entity) might well, from a different perspective, turn out to be internally structured. For instance, a corporation or a university might function and interact as if it were an individual, for example, by entering into contracts with others. Yet, from an internal perspective, it would generally consist of several agents.

Combining the above intuition with the context-based view of commitments is that it enables us to express complex domain structures in a simple manner. For instance, we can imagine a team as the organizational context of the several commitments that tie together its members. However, the team is itself constructed from its members. Thus we would naturally model commitments between the team viewed as an agent and each of its members. Such commitments might capture the principal intuitions of teamwork such as that a member of a team should support the other members of the team in succeeding with their goals, and that the team-members should coordinate with one another to accomplish their common goals. The specification of commitments between the members and the team help codify such relationships. The team-members need not form mutual beliefs or joint intentions with one another, as traditional approaches require [30, 23], because the essence of the relationship between them can be captured through the commitments. In particular, a team-member may not even know who the other team-members are in order to form a social relationship with them through the common identity of the team to which they belong. A further benefit is that the relationships naturally express the rules of encounter of the team and thus support the expectations that the team-members might form on each other. Additionally, the relationship can potentially be realized in a variety of ways. For example, the members of a team may join it one by one and a team-member may leave and another may join without altering the essential fabric of the team. Or, the team-members may all join at once.

An important aspect of commitments is that they can be manipulated [40]. A debtor may create or cancel a commitment; a creditor may release it. More interestingly, a debtor may delegate a commitment to a new debtor, and a creditor may assign it to a new creditor. Such manipulations provide a high-

level and systematic way in which the social state can progress. Fornara and Colombetti [19], [20] have studied the operationalization of commitments to support such manipulations. With Xing [52] and Chopra and Desai [46], I have further developed patterns involving the manipulation of commitments that support useful properties.

This wasn't always emphasized in the early works, but the conditionality of commitments is important. By default, commitments are conditional, involving an antecedent and a consequent, and unconditional commitments are merely the case where the antecedent is true. In logical terms, the conditionality of commitments resembles that of a strong conditional rather than a material conditional [43]. A commitment becomes and stays *detached* or *discharged* when, respectively, its antecedent and consequent become true. There is no presumption of temporal order between the detach and discharge of a commitment. A commitment that is detached but fails to discharge indicates a violation.

It is worth distinguishing two major kinds of commitments. *Practical commitments*–as commonly seen in formalizations of business processes–are about what the debtor would bring about. *Dialectical commitments*–as commonly seen in formalizations of argumentation–are about what the debtor stakes a claim on. The import of the two kinds of commitments is quite different and parallels the distinction between goals and beliefs, respectively. Practical commitments call for action and thus relate to present or future actions. Dialectical commitments call for a condition holding and thus can relate to past, present, or future.

That the two kinds of commitment are distinct has been known for years and, in particular, finds discussion in some of Professor Castelfranchi's work wherein he provides the clearest exposition of it. However, the distinction seems to have been lost in the agents literature until recently. I have sought to revive this distinction in conjunction with a proposed formal semantics for commitments [43].

3 What are Commitments Good For?

In a nutshell, commitments form a key element, arguably the most important element, of the social state of two or more interacting agents.

Commitments are important because they help us address the tradeoffs between and reconcile the tension between autonomy and interdependence. On the one hand, we would like to model our agents as being autonomous with respect to each other. On the other hand, it is clear that if the agents were fully autonomous, then we would have not a multiagent system in the true sense of the term, but merely a number of agents that happen to coexist in a shared environment. Such a system would exhibit no useful structure. Further, it is clear that autonomous agents must be able to cooperate and compete with each other, and carry out complex interactions. If there were no interdependence, the agents would be nearly useless. Professor Castelfranchi and his colleagues

first articulated the importance of such interdependence among agents and explored varieties of it [36]. Similar intuitions and elaborating the connection with autonomy arise in newer work [27]. Commitments provide a natural way to characterize the bounds of autonomy and interdependence without getting bogged down in low-level details.

3.1 Commitments for Business Protocols

A business protocol characterizes how a family of interactions involving two or more business partners may proceed. What makes a business protocol "business" is that the interactions it characterizes involve business relationships. The classic examples of business protocols are those realized in cross-organizational business processes, such as for negotiation, sales and purchase, outsourcing of various business functions, delivery, repair, and so on. Traditionally, business protocols have been modeled in purely operational terms such as through state transition diagrams or message sequence charts that describe ordering and occurrence constraints on the messages exchanged, but not the meanings of such messages.

Commitments provide a natural basis for capturing the meanings of the messages. In this manner, they provide a standard of correctness. A participant in a business protocol complies with the protocol if it ensures that if any commitment (of which it is the debtor) is detached, then it is also discharged (that is, not violated or canceled–neglecting the distinction between them). Having such a declarative basis for correctness not only simplifies the modeling of the interactions being designed or analyzed but also provides a basis for flexible enactment that can be shown to be correct.

A typical use of commitments in business protocols involves introducing the syntax for the messages under consideration along with a formalization of the meanings of the messages expressed in terms of the commitments of the participants and the domain or environmental propositions that have a bearing on those commitments. For example, in a purchase order protocol, we might introduce a message *offer* and define its meaning as involving the creation of a commitment–with its sender being the debtor and its receiver being the creditor. The commitment would specify that the sender would provide the goods to the receiver were the receiver to accept the terms. Likewise, we might introduce a message *accept* through which the recipient of the *offer* would take up the given offer. Based on these meanings, we would be able to determine if an enactment of the protocol was sound. Even a simple and obvious specification of correctness proves effective: this states that an agent complies with a protocol if no enactment of the protocol ends with the agent as the debtor of a detached but not discharged commitment.

The natural connection between commitments and correct enactments naturally leads to ways of operationalizing them. Each commitment provides a basis for judging the compliance of its debtor. The commitments of interest taken together provide a *public* or neutral perspective on the correctness of an

interaction as a whole. Further, the idea of using both commitments that refer to the antecedents and consequents of other commitments and commitments that refer to the creation and manipulation of other commitments provides a powerful basis for capturing a network of social relationships at a high level. An agent can thus reason about the commitments of interest to it, especially those where it is the debtor or creditor, and decide how to interact with the other agents participating in the current business protocol. Although the agent may act as it pleases, the commitments themselves impose constraints in terms of what actions are compliant. In this sense, the specification of commitments leads to the notion of a commitment machine [10, 51, 53, 42].

When we apply commitments as a basis for the semantics of the communications among agents, they yield a basis that is formal, declarative, verifiable, and meaningful [41]. Interestingly, commitments also lend themselves to operationalization in a more traditional manner. This is the idea of compiling a commitment machine into a traditional representation such as a finite state machine over finite [10, 51, 53] or infinite [42] computations. Such compilation removes the opportunities for flexibility that an explicit commitment representation supports. However, a finite state machine can be executed by agents who are not equipped with an ability to reason logically. Moreover, such a mechanically produced finite state machine can often be more complete in its coverage of important scenarios than a hand-generated one–and consequently be too large and unwieldy for a human designer to specify by hand.

3.2 Commitments for Communication Languages

The above idea involving protocols can also potentially be applied as a basis for the meanings of the primitives in agent communication languages (ACLs). ACL primitives have traditionally been given semantics based on the beliefs and intentions of the communicating parties. Instead, a commitment-based semantics could naturally express the social relationships between the communicating parties. In essence, one would take the idea of commitments for individual communication protocols and apply that idea to the modeling of general-purpose communication primitives. The idea is not implausible in itself. It is indeed possible to define the meanings of communication primitives. In spirit, this is not different from the meanings of the messages in the business protocols. However, the particular formulations in this setting suggest ways to capture richer subtleties of meaning than may be necessary in a typical business setting. In particular, I have suggested [41] that meanings can be captured via a trio of specifications that, following Jürgen Habermas [24], reflect objective, subjective, and practical meanings. These types of meaning can be expressed in terms of commitments regarding, respectively, the relevant aspects of objective, subjective (cognitive), and practical (subjective and institutional, with an emphasis on the latter) reality.

For example, we might define an *informative* message type as one creating a dialectical commitment with its sender as debtor, its receiver as creditor, its

antecedent as true, and its consequent as asserting the truth of the proposition specified as the content of the message. In the above terms, this would be the objective meaning [41]. Likewise, a *commissive* message type would create a practical commitment. And, similarly for the rest. I should note in passing that the idea of a general-purpose ACL itself is suspect (see my recent manifesto, as included in [15], for a discussion of this point). In any case, we can view the definitions of the primitives as useful patterns, which might be specialized and applied to the communicative acts needed for particular protocols.

3.3 Commitments and Conventions

A deeper benefit of commitments is in their relationship to conventions. Two levels of abstraction are worth distinguishing in formalizing even the simplest interactions. First, a quote means that there is a commitment from the merchant to sell the specified goods at the specified price. Second, the fact that the quote means the above is a matter of convention in the chosen domain of commerce, and therefore both the merchant and the customer commit to that meaning. Specifically, the meanings of any communications must be based on the conventions at play in the given social setting. It is thus highly natural that we understand conventions as a basis for interactions among autonomous parties.

In several cases, the applicable conventions would be determined based on longstanding tradition in a domain; in other cases, they may be explicitly negotiated. For example, in the financial capital markets, a price quote for a stock (sent by a broker to a trader) is interpreted as being merely informative of the last known price at which that stock was traded. In typical commerce, however, a price quote (sent by a merchant to a customer) can be interpreted as an offer to sell at the specified price. In the latter case, the longevity of the offer can vary: for a business-to-business supply price quote, the offer may be valid for 30 days whereas for an airline to consumer ticket price, the offer may be valid for a minute. The longevity of the offer too is often a matter of convention. The importance of conventions to meaning and interoperation among autonomous parties is thus quite obvious.

What is interesting for us is that the conventions that arise in a given setting can be expressed as commitments. Specifically, each of the parties involved (or sufficiently many of them) would commit dialectically to the existence of the convention. Dialectical commitments, as are involved in this case, are different from the practical commitments involved in formalizing the messages in typical business protocols. However, each party may additionally practically commit to acting according to the conventions. Arguably, something prevails as a convention in a community only if the participants dialectically commit to it and practically to acting according to it.

The general notion of conventions and especially as related to agent communications [28], however, merits study in its own right. The interesting observation from the standpoint of commitments is that a convention corresponds

to an aggregation of dialectical commitments. The commitments can be structured using the context as explained above. Thus the participants in a community where a convention prevails can dialectically commit to the convention. Each participant would be a debtor and each other participant would be a creditor. Alternatively, the creditor could be the context and thus stand for the community as a whole.

4 Concomitant Confusions

In the worlds of artificial intelligence and software engineering abstractions, commitments are the new kid on the block. A common prejudice in these traditional disciplines that finds its way into multiagent systems is to formulate the design problem as one for a complete unitary system, even when such a system is to serve the needs of multiple stakeholders. Hence, all too often, researchers and practitioners approach the design of a multiagent system not only as consisting of cooperative (and sincere) agents, but also as one where they will themselves provide all of the agents.

In contrast, commitments are most germane and offer their greatest value in settings where capturing the meanings of the interactions being designed is relevant. We would leave the design and construction of the agents to their implementers even though in some cases we might ourselves take on the implementation task. Further, we would leave the operation of the agents to the agents and their users. That is, commitments can apply in traditional settings where all agents may be designed by one party, and can help specify cleaner architectures. But they are not confined to such settings, and the assumptions needed for a unitary system do not apply in general to commitments.

One can imagine an engineer thinking "well, I am going to design a good system of three agents; I am going to make sure the agents take on goals and beliefs that are compatible with their commitments and adopt policies that help them realize their commitments; and I am going to damn well make sure the agents walk the straight and narrow, so I will prevent them from violating their commitments." Such thoughts may well be appropriate in a single-perspective, cooperative, regimented system constructed by one engineer from a set of agents. I would place the work of Minsky and Ungureanu [33] into this category who are not focused on cognitive agents but on conventional architectures, in which setting their approach is more reasonable. However, such thinking unnecessarily limits the multiagent systems designs that one comes up with. Therefore, although such thinking may be a useful design pattern to help relate open architectures to traditional architectures, when framed as a general constraint on commitments, it is misguided.

In simple terms, we can separate three scopes of effort or decision making: (1) the modeler of an interaction defines interactions via their associated commitments; (2) the agent designer implements an agent; and (3) the agent (and its users) decide how to behave on the field. The multiagent system engineer must specify the interaction precisely and relinquish control of the design and

operation of the endpoints of the interaction. Relinquishing control is a consequence of dealing with open systems. Focusing on interactions is the only plausible way of engineering a system where the engineer lacks control over the endpoints.

4.1 Commitments versus Goals

A common view is that an agent who commits as debtor to bringing about a condition in the world also adopts the same condition as a goal. (In some accounts, the agents would adopt an intention, not just a goal, but let us disregard the distinction between goals and intentions here.) A stronger variant is when the goal applies to both the debtor and the creditor of a commitment. This confusion is insidious because it relies upon a careless reading of the literature: the confusion is nothing more than a confusion between the S-commitments (our commitments here) and the P-commitments (traditional commitments in AI planning) as explained in Section 2.

Commitments and goals are fundamentally different kinds of creatures. A commitment is a public or observable relationship between two parties whereas a goal is a single-agent representation. An agent's commitments are generally known to others because of the conventions in play in the given setting. An agent's goals are never inherently known to another agent, although another agent might reason about them based on assumptions of rationality or based on explicit revelation by the first agent, provided appropriate conventions apply to the presumed revelation.

It is true that in general a cooperative debtor that created a commitment would simultaneously adopt the corresponding goal. However, an agent may not adopt the corresponding goal, potentially risking failing with its commitment–and thus risking harm to its reputation and risking additional sanctions of penalties and censure. Conversely, an agent may hold a goal and not have committed to any other party for it. Such a goal might well be a highly important goal for the agent–after all, a goal would relate to the agent's preferences, and not necessarily to something the agent would reveal to others.

As a telling example, consider the common situation where an airline operating a 100-seat airplane books 120 passengers on it. Clearly, the airline is committed to each ticketed passenger, but equally clearly the airline could not have a goal to board each passenger on to the airplane. The airline simply has a clever internal strategy to maximize profit where it knowingly enters into commitments that it might not be able to discharge. If 20 passengers miss the flight, the airline goes scot-free; if more than 100 show up, the airline compensates those it does not take on board, but it still comes out ahead on average.

Misalignments between commitments and goals are not the same as deception. In the above example, the airline has no intention of misleading its clients. In fact, the airline may strongly believe–based on the evidence at hand–that no more than 100 passengers will show up and thus none of its commitments would be violated. However, it is fair to say (as a reviewer suggests) that a

commitment that is supported by its debtor's goal would be likelier to be effective provided the debtor is sufficiently competent.

4.2 Commitments versus Beliefs

It is not uncommon to conflate commitments with beliefs. The motivation seems to be that an agent would represent its commitments and thus believe them to exist. But such an argument would hold for just about any representation.

In some cases, there is a more subtle confusion between commitments of the dialectical flavor and beliefs. Notice that even dialectical commitments are commitments, meaning that they reflect their debtor staking a claim or accepting a claim as a putative fact such as, for the sake of discussion, during an argument [32]. The debtor may not in fact believe what it commits to. Conversely, the debtor may have numerous beliefs it keeps private and never commits to holding to another agent. Any such commitment binds the debtor to a certain pressure to interact in a certain way, and there is generally little reason to expose all beliefs as commitments.

4.3 Commitments versus Mutual Beliefs

A more insidious confusion arises with respect to mutual beliefs. As Section 2 explains, the underlying idea behind using the mutual beliefs (and equally intentions) was to introduce a level of mutuality while continuing to use the mental concepts.

The first problem with this view is that it is wrong. Commitments are not mutual beliefs. A commitment is a unidirectional relationship. For example, if Bianca offers to sell a camera to Alessia, the commitment holds whether or not Bianca believes it or Alessia believes it. As in the airline over-booking example, Bianca may simply have made the offer to try to prevent Alessia from taking up another offer. And Alessia might be on to Bianca: that is, she might not believe that Bianca believes she would supply the camera. However, the commitment exists. Alessia may in fact file a complaint against Bianca. Alessia would not be able to file a complaint if the commitment was defined as the mutual belief.

The second problem is that mutual beliefs are extremely fragile. Let us say Alessia believes that Bianca believes that Alessia believes . . . that Bianca will be shipping a camera to Alessia. If Alessia believes that, at the hundredth level of the nesting, Bianca might not believe Alessia expects the camera any more, that would dissolve the mutual belief. However, in real life the commitment does not go away in such a case. Bianca is not off the hook based on a failure of a belief by Alessia and certainly not for imagining that Alessia may have lapsed in her belief.

The third problem is that, again at variance from real-life interactions, although commitments arise in all manner of distributed settings, mutual beliefs generally cannot be constructed. Under asynchronous communication, the only mutual beliefs in a system are the invariants of the system, that is, propositions that were true from the start [8]. Indeed, the artifact of mutual beliefs (along with the similar artifact of common knowledge) is used in distributed computing primarily to prove *impossibility* results [8, 25]. Because mutual beliefs cannot be engendered through message exchange in general asynchronous settings, a problem that requires mutual belief is unsolvable.

Clearly, the AI researchers have understood the problem in terms of live human communication, which is inherently synchronous. In multiagent settings, they address the challenges of asynchrony by fiat. Specifically, they assume that a single message by one party to another, without any need for an acknowledgment, would achieve mutual belief. The idea it seems was that there was a central belief store and any assertions inserted into it reflected the beliefs and further even the mutual beliefs of everyone in the system. However, AI researchers by and large hide this key assumption in the implementations of their systems and never mention it in their theoretical descriptions.

One could treat the above assumption (of a single message exchange being sufficient) as a standard operating procedure, as mentioned in Section 2, in a particular setting. But that only means we are seeking to characterize commitments a certain way. So why not be honest and model the commitments directly? About the only reason not to do so is if one has locked on to the mentalist ideology.

I should explain that the point is more general than merely one of physical transmittal of information, as it is in the traditional distributed computing literature. The deeper and more crucial point is of the necessity of simultaneously sustaining multiple perspectives. In other words, what is most problematic is not so much the physically central nature of the belief store where mutual beliefs might exist, but its conceptually central nature, indicating that we had magically consolidated the perspectives of multiple autonomous, heterogeneous parties into a correct unitary perspective.

4.4 Commitments versus Obligations

Obligations are an important notion studied since ancient times. A traditional obligation applies on an agent, roughly corresponding to the debtor of a commitment. What distinguishes a traditional obligation from the cognitive concepts of beliefs and goals is that it is inherently externally focused: an obligation can be met or not and the consequences occur beyond the minds of the agents involved. A more interesting kind of obligation is directed: here an agent is obliged to another agent [26]. The second agent corresponds to the creditor of a commitment.

Because directed obligations are clearly interagent in their orientation, they are a more natural match for multiagent systems than are traditional obli-

gations. The similarities between directed obligations and commitments are striking. But can we treat commitments as being identical to obligations? A commitment when it is active corresponds to a directed obligation.

However, commitments and obligations have important points of distinction. First, a commitment can be manipulated, in particular, delegated, assigned, or released. Second, a commitment carries with it an organizational context, as explained above. Third, obligations carry a moral connotation that commitments lack. Fourth, a commitment reflects the inherent autonomy of the participants in an interaction. Thus an agent would become a debtor of a commitment based on the agent's own communications: either by directly saying something or having another agent communicate something in conjunction with a prior communication of the debtor. That is, there is a causal path from the establishment of a commitment to prior communications by the debtor of that commitment. Obligations by contrast can be designed in or inserted by fiat.

Frank Dignum observes (in a private comment) that the autonomous nature of commitments raises a creditor's expectation that the debtor's goals and beliefs are aligned with the commitment, and hence it should be discharged. This point applies to cooperative debtors and may be a basis for the conventional interpretation of communications in general.

4.5 Commitments versus Policies

A commitment, especially in its conditional form, looks like a rule for processing, and in this sense resembles a policy. For example, an engineer might take the view that an offer from a merchant to a customer expresses the merchant's policy that if the customer pays the specified amount to the merchant the merchant will send a cello string of a specified type to the customer.

Treating a commitment as a policy in this sense reflects the same confusion as with goals and beliefs, namely, that the external, interactive, observable nature of commitments is conflated with the internal, behavioral, private nature of another abstraction. A policy is how an agent may decide to act upon–or decide not to act upon–a commitment. If the merchant has a straightforward policy for acting on all its commitments, then so much the better.

However, note that in general, a commitment would be necessarily incomplete with respect to the behavior needed to discharge it, and thus the policies associated with a commitment may need to specify aspects that the commitment does not mention. In our example, the merchant would have committed simply to supplying, say, a Larsen cello D string for payment. The merchant would need additional policies to determine how exactly to supply the D string. Should the merchant supply the instance of the D string that is the oldest in its inventory? Or, the newest? Or, one that happens to be the most convenient based on other tasks the merchant is performing, for example, supply from the top rack if the ladder is up there anyway, else supply from the bottom rack? Maybe the merchant will do well to supply a carefully checked

instance of the string to a repeat customer and supply it with extra robust packaging for a customer overseas or for a customer who has a high standing in the user community and can influence other prospective customers.

These are all legitimate policies, but it would be inappropriate to tie them into the commitments. Indeed, were we to attempt to specify commitments at the level of such policies, we would face important challenges and produce a poorer quality model as a result. The challenges would be first coming up with detailed specifications and second, importantly, finding a way to determine if a party is complying with the commitment–for example, how will we ever know if the merchant sold the oldest item in its inventory? Or, one from the top rack? The resulting model would be of poor quality because it would tightly couple the parties involved in the interaction. For example, a merchant who committed to supplying the oldest item (and did so honestly) would be compelled to maintain information about the ages of the items in its inventory and to set up its internal business processes to search for items in their order of age. It would not be able to take advantage of any improvements in internal business processes as might arise later. Equally importantly, a new merchant who wished to join the interactions specified by such a commitment would not be able to participate without developing such otherwise irrelevant components of its information systems.

There is another notion of policies, however, which does make sense when related to commitments. This is the idea of a social policy, which captures the rules of encounter in a society. I have occasionally used the term "social policy" in this sense, but I now think it is better to refer to such as *norms* and to reserve the word "policies" for the policies of an agent or organization that reflect its decision making.

4.6 Commitments versus Regimentation

I encounter this problem a lot in discussions with conventional software engineers. They are accustomed to capturing requirements for, modeling, designing, and implementing software systems in which there is a single locus of autonomy. The system in question may be distributed but it is conceptually unitary and involves the perspective of a single party. You can identify such a mindset where the engineer talks of "the system" as an entity that will interact with "the user"–the goal of the engineer is to create a software design artifact from which one can develop a set of software modules that will meet the elicited requirements as the system interacts with its users.

In such a case, when the engineer begins reluctantly to think of social interactions and commitments among the parties involved, the engineer's mindset remains to try to force the modules to behave in the "correct" manner. The engineer attempts to capture such behaviors via commitments. In other words, the engineer retains the single-party perspective and, without absorbing the idea of any social interaction among notionally autonomous parties, merely treats commitments as a clever-sounding representational framework.

The engineer's challenge is to force the modules to adopt certain commitments and to act on their commitments in exactly the chosen "correct" way. All too often, such designs emerge from when a traditionally minded software developer reverse-engineers an existing process into the representation of commitments–adopting and incorporating every ad hoc quirk of the original model into the commitment-based model.

Following Jones and Sergot [29], I term such a viewpoint *regimentation*. In general, the use of regimentation obviates the need for modeling commitments. However, for engineers new to commitments, it might be a useful intermediate step provided they recognize it as such and proceed to develop an interaction-oriented model.

4.7 Commitments and Compliance

When computer scientists and business (process) modelers first encounter commitments, they immediately ask about compliance: how can we guarantee that an agent would comply with its commitments, or at least not wantonly violate or cancel them? For a novice, this question is reasonable. But upon reflection, we can see this question is misguided and unfair because it hides some crucial presuppositions and confusions. Underlying this question is the misguided assumption that if one simply fails to model–or even acknowledge the existence of–an agent's commitments, the agent would behave perfectly.

A strange variation on this theme is that if we were to model communications among agents in terms of commitments we would have created legal liabilities that didn't exist before. No, seriously, I am not making this up. The idea is that if Bianca sends Alessia a message with an offer for a camera, for example, using English or XML, it is just fine and legally safe. But if we only so much as realize that the offer is a commitment to provide the specified camera, Bianca would become liable in ways that she wasn't when we didn't model her English or XML message as conveying a commitment. Perhaps the people who come up with the above variants imagine that obfuscation of meaning is a legal defense. I claim, instead, that for a business or other interaction to be successful, the parties involved must share an understanding of the terms involved. In general, even lawyers would prefer greater clarity as a way to define each party's expectations of the others.

Modeling commitments does not cause agents to (potentially) behave in an undesirable manner. Indeed, modeling commitments helps potentially address the challenge of ensuring compliance. By treating commitments explicitly, we (1) obtain a crisp, yet not operational, statement of compliance; (2) formulate the notion of *transparent* protocols in which compliance determination is possible; and (3) open the way for designing agents using beliefs and goals who will be compliant with their protocols. Monitoring and compliance relate naturally to themes such as formalizing (1) organizations and governance [50, 21], for example, penalizing malfeasant agents in a community, and (2) bases for relating commitments and economic models of rationality [16].

4.8 Terminological Confusions

It is worth highlighting here some confusions that arise alarmingly frequently, though usually among people who are unfamiliar with the commitments literature. At the root of these confusions are lexical mismatches, wherein the reader misinterprets a technical term, even though the terms under consideration are well-defined in the commitments literature.

Commitments are psychological. The raison d'être for commitments is to avoid the shortcomings of psychological commitments, but that doesn't stop some people from inadvertently going back to square one.

Social means going to a bar. We use the term social to distinguish from psychological, not that commitments are only about cultural conditions or for after-hours socialization. The most common application of commitments today is in modeling business organizations and interactions, though there is no reason to preclude other settings even personal relationships.

Private means shared. Private refers to the internals of an agent and public to what is shared or observable. If one agent commits to another, that means we have created a social object involving at least two agents. Even if the two agents keep the commitment confidential, never disclosing it to a third party, the fact that it involves more than one agent makes it public, as we define the term.

Debts are exclusively financial. We simply use debtor and creditor to indicate the directionality of commitments. These terms are reminiscent of their usage in the vernacular, but generalize over it. There is no restriction to financial debt: the conditions involved could be arbitrary; indeed, even in normal English, debts are not restricted to be just financial.

Organizational context means any element of the situation. But organizational context is *not* just anything: in our technical meaning, it is an objective institutional construct treated on par with an agent.

Commitments are ontological commitments. In Quine's [34] terminology, an ontological commitment describes the objects one entertains as existing. For example, if I say my grandfather owned a unicorn, that means I am ontologically committed to the past existence of at least one unicorn (and of my grandfather, and of the two existing contemporaneously). Ontological commitments resemble presuppositions underlying utterances that a person makes whereas commitments for us are about actions or staked claims. One could formulate a dialectical commitment for the existence of anything that its debtor makes an ontological commitment to.

5 Debate with Professor Castelfranchi

Let me now turn to the most interesting part of this article, which is to highlight some of the points of difference between Professor Castelfranchi's views and mine regarding commitments.

Let me begin with a point, which I think is not controversial, though potentially sounding like it might be. At the root of it is the emphasis I place on the importance of observable interactions among agents (including low-level behaviors), which contrasts with Professor Castelfranchi's emphasis on the cognitive representations of the agents. I suspect unclarity on my part led Professor Castelfranchi to criticize my approach as resembling a behaviorist approach.

A behaviorist stance would be reduced to entertaining nothing beyond (what the designer or analyst imagines are) the objective atoms of behavior. In general, the difficulty in identifying such objective atoms is indeed one of the challenges that uproots behaviorism. The acute need for imagining what is ostensibly objective is one of the shortcomings of behaviorism. However, I do not see that commitments can be reduced merely to low-level behaviors. Instead, here we are accommodating a rich social reality: we have postulated agents who create and function in social institutions, who entertain abstract high-level relationships such as those expressed via commitments, and who not only communicate at the level of exchanging bits of information but also communicate in suitable institutional terms.

Professor Castelfranchi and I thus agree that the study of commitments is not and should not be treated as a behaviorist project. Instead, our collective effort in multiagent systems may be thought of as a realist project in that we treat common-sense social constructs such as commitments as real entities.

5.1 Commitments and Autonomy

Broadly speaking, the multiagent systems field is primarily concerned with understanding the interactions of agents. At a basic level the autonomy of the agents is key. Of course, fully autonomous agents would be useless if not harmful–clearly, what we need to understand is the interdependence of the agents. That is exactly where commitments come in. Each commitment captures one element of a social relationship between two parties. When we put these elements together, we obtain the network of relationships that characterizes a multiagent system. I expect that Professor Castelfranchi and I agree on the above in broad terms.

Where I suspect we disagree is in the relative importance we accord the intuitions of autonomy and interdependence. As I see it, an agent must be able to enter into and exit its commitments at will whereas Professor Castelfranchi sees the process as more constrained. These distinctions become more apparent when we consider the creation or cancellation of a commitment.

5.1.1 Accepting a Commitment

Professor Castelfranchi sees a commitment in a positive light whereas I see it as a general notion in a neutral light. Also, my interest is to maximize the

flexibility of the interactions and the autonomy of the participants. As a result, I would consider a commitment to be created if its debtor says so. In this sense, the creation of a commitment is a declarative or performative communication and is within the control of the agent initiating that communication, given the appropriate circumstances and conventions. In contrast, Professor Castelfranchi would like to see the creditor of a commitment explicitly accept the commitment before it comes into being.

A downside to Professor Castelfranchi's approach is that it couples the two agents unnecessarily. It also differs from common uses of commitments. For example, a merchant can make an offer to a customer merely by saying so. The customer may sit silently for a while (up to the time period of the offer) and then attempt to make a purchase based on that offer. That is, the customer doesn't separately accept the offer and then exercise it; the customer simply exercises the offer directly. The offer is valid all along. If we were to require that the offer be accepted before it comes into existence, that would seem to require that a message exchange has to complete before the offer begins to exist.

Professor Castelfranchi is concerned that if we do not include an explicit acceptance, an agent may in essence use a commitment to make a threat, for example, by committing to harm the creditor. In Professor Castelfranchi's approach, the creditor would refuse such a commitment and thus never let it be formed. Notice, however, a malicious (prospective) debtor could harm the creditor nevertheless. If the commitment happens to be undesirable for the creditor, it could (i) resist it in other ways, perhaps by making a threat of its own; (ii) ignore the commitment and not demand that the debtor discharge it; (iii) assume it arose due to some underlying confusion due to miscommunication with the debtor, and explicitly release the debtor from that commitment. Each of these sample approaches has the advantage of not creating avoidable coupling between the debtor and the creditor.

Also, the apparent undesirable-to-the-creditor orientation of the content of a commitment cannot always be avoided. For example, an organization's president Alessia may have committed to all its members that she would punish the treasurer were the treasurer to embezzle any funds. A member, Bob, may accept the commitment at a meeting along with the other members of the organization. Now later if Bob becomes the treasurer, he would be the creditor of a commitment from the president that might potentially penalize him, if it is activated at all.

An alternative view is that the above notion of acceptance ought to be considered as being explicit *or* implicit. Thus silence in our example above can be treated implicit consent. This view, however, misses two important points. The first point is that it contravenes the agents' autonomy, as explained above. The second point we can explain as follows. The deeper purpose of talking about commitments is to help us understand the social state of an interaction. If we decide that a commitment is created only upon acceptance by the prospective creditor that means we can provide no clear meaning for the intermediate state wherein the debtor has "committed" but not quite because the creditor has not confirmed yet. If we allow implicit acceptance, then we have no viable basis for distinguishing between the commitment and its half-

baked stage. That half-baked commitment is not nothing because the debtor is on the line if the creditor accepts it. I claim that if the associated intermediate social state were to be formalized properly, the semantics that results from the acceptance-based approach would be close to that of the one-sided formulation that I advocate.

Consider the following example, which came up in a discussion with Neil Yorke-Smith. How might one model the following? Alessia proposes to Bob that they exchange goods for payment tomorrow, but today Alessia would like to know whether Bob accepts or not.

A simple formulation is C(Alessia, Bob, C(Bob, Alessia, goods, pay), goods), indicating that Alessia tells Bob "if you commit to pay on receipt, I will send you the goods." It's Alessia's decision to trust Bob. If Bob does commit, Alessia must send the goods or violate her (now detached) commitment. If Alessia sends the goods after Bob's acceptance, Bob must pay or violate his (now detached) commitment. This formulation shows how we can make the acceptance of a commitment explicit if and when we need it to model some scenario, but do not need to insist upon acceptance in other cases. We can think of the above formulation as interpolating two one-sided commitments Alessia to Bob: one conditional on payment, C(Alessia, Bob, pay, goods) and the other unconditional C(Alessia, Bob, true, goods). In contrast, the acceptance-based representation makes it impossible to express the one-side commitments; tends to be applied wrongly wherein one agent commits another, thereby violating the latter's autonomy; and, leaves as undefined the social state wherein Alessia has made an offer but Bob hasn't responded.

5.1.2 Accepting a Cancellation

In much the same spirit, I propose that an agent can cancel its commitment at will. Like creation, a cancellation is a declarative that the debtor can perform. Likewise, a creditor can perform the release of a commitment at will. In the case of the cancellation, the outcome might not be one that the creditor desires or would willingly accept; further, the outcome might be one that we as designers might not condone in our agents. However, if that were to be the case, the creditor should have made sure (or we, the designers, should have made sure) that there will be repercussions on the debtor for having performed an inappropriate cancellation.

One might think that these repercussions signal the unacceptability of the cancellation and, therefore, that cancellations should only be allowed when the creditor accepts. I won't repeat the points made above in connection with creating a commitment, which apply here too. However, an additional point relevant to cancellation is that in a multiagent system (consisting of autonomous agents), we can rely on *regulation* but not on *regimentation* [29]. Regulation is about controlling behavior through normative means whereas regimentation is simply about preventing bad behavior [1]. Regulation is suited to interactions among autonomous agents. In contrast, regimentation–which here

corresponds to preventing cancellation by explicit acceptance–contravenes autonomy.

Even in the original formulation of commitments [38], the notion of the (organizational) context of a commitment served to accommodate such cases. Specifically, if the cancellation of a commitment arises because of true and reasonable exceptions, the context may impose no penalty upon the debtor; in other cases it might. For example, let's say a merchant has committed to providing some goods to a customer. If the merchant cancels the commitment to do so because of a tsunami that destroyed the manufacturing plant and refunds the customer's payment, the cancellation appears not unfair whereas if the merchant cancels because the merchant can now demand a higher price, the cancellation does sound egregious. Let us say the (organizational) context here is the electronic marketplace, for example, eBay. In the first case, the context may declare the cancellation legitimate; in the second case, not so. In the second case, the context may penalize the merchant, for example, by revoking his credentials in the marketplace or pursuing fraud charges in the court system.

If the organizational context can ensure such coherent outcomes, then we can think of the context (and the concomitant family of interactions) as being well-designed (notice we make no claims about the internals of the agents themselves). If the context is not well-designed, then either we as designers made a mistake or the agent (customer) made a mistake in joining such a context, dealing with an untrustworthy merchant, and foolishly counting on him to discharge his commitments.

5.2 Commitments and Cognition

Another of the points where I continue to have a disagreement with Professor Castelfranchi is in the function and importance of cognitive representations in connection with commitments. We agree, of course, on the basic idea that an agent's behavior is of central importance in judging whether or not it discharges its commitments. And, I expect we agree not only on the essential relevance of commitments to the social life of an agent, including its relationships with other agents, but also on the importance of cognition.

Professor Castelfranchi, however, assigns a far stronger function to the cognitive representations of an agent than I do. To him, having a commitment is strongly based on the associated patterns of beliefs, goals, and intentions. For me, in contrast, a commitment is a social entity, which takes its existence from the public sphere. An intelligent agent would undoubtedly represent and reason about its commitments, and its commitments would undoubtedly affect and be affected by its goals and intentions. However, to my thinking, a commitment at its core remains purely social. In this regard, a commitment is no more and no less of an abstract object than any cognitive attitude or any mathematical object for that matter–that is, a commitment can exist in the public sphere just as legitimately as in the mind of an agent.

Although I recognize the benefits and importance of the cognitive representations in modeling and implementing agents, I consider such representations to be internal to an agent and reflective of its internal architecture and construction. In contrast, I understand commitments as having normative force whereby they can provide a potentially independent basis for judging the felicity and correctness of the actions of agents. When we define commitments in such a public and observable manner, they can become a key ingredient in understanding the institutional nature of communications and indeed of understanding institutions themselves.

As an example, consider a friend of mine who promises to help by giving me a ride to the airport. My friend would have done so by using the prevailing vernacular of our social institutions to create a promise. Let us say the appointed hour comes and goes, but my friend does not materialize. Thus he has violated his commitment. For the sake of this example, let us further stipulate both that I trust my friend in such matters and that he is highly trustworthy in fact and would not have deceived anyone. Clearly, he forgot or found himself in a personal emergency. But we would still state that he violated his commitment, albeit inadvertently or in exonerating circumstances.

We should be able to pass the judgment of the commitment being violated based on what we observe, namely, the failure of the commitment. However, if the definition of commitments were to be intertwined with questions of beliefs and goals, it would be difficult for us to pass even such elementary judgments. Further, the definition would lose the benefit of modularity by combining the social and the cognitive representations. Additionally, it would create a situation where we would not be able to determine if a commitment existed without being able to assess what the beliefs and intentions of the parties involved were, and it is well-known that such ascriptions cannot be defended in multiagent settings where the agents are not homogeneous and their internal states not public [39].

I claim that such judgments provide the basis of the normative strength that commitments carry. We might conduct any amount of elaborate post mortem analyses involving the beliefs and goals of the participants, but if we are not clear about the objective fact in this matter, we lose not only a basis for specifying an institutional basis for multiagent systems but also for conducting any cognitive analyses with any grounding in truth.

6 Themes for the Future

6.1 Commitments and Trust in Social Computing

The increasing attention garnered by topics such as social computing tells us that areas of long interest in the multiagent systems field [22] and especially pursued by Professor Castelfranchi himself [5] are gaining currency. Today's practice in social computing is weak indeed and consists of little more than users sharing information on a social networking site or users performing

various assigned tasks in what is called crowdsourcing. It seems to me self-evident that any kind of realistic social computing must rely upon the concepts of commitments and trust.

The study of trust has been an important theme in Professor Castelfranchi's body of research. Professor Castelfranchi and colleagues have developed a semantically rich notion of trust [6, 7, 18] that incorporates both its social and its cognitive aspects. Professor Castelfranchi's approach contrasts with the majority of computer science works on trust, which tend to jump into (typically, numerical) representations without first sorting out what the trust as conceived stands for. Professor Castelfranchi relates trust to the plans of the parties involved and their expectations with respect to each other. I find another of previous works by Professor Castelfranchi and colleagues as especially germane here. This is the notion of dependence [36], which Rino Falcone and Professor Castelfranchi [17] have recently revived and related to trust.

It seems clear to me that these concepts suggest the strong relationship between commitments and trust. In conceptual terms, we can think of commitments and trust as duals of each other: a debtor commits to a creditor and a truster places trust in a trustee. The idea of commitments as expectations in reverse originates in Amit Chopra's [11] dissertation. I have recently begun to formalize trust in a manner that highlights the notion of dependence and relates trust to commitments [44]. Not every commitment may have corresponding trust in the reverse direction. And, not every placement of trust may be justified by a commitment in the reverse direction. The best outcomes arise when trust and commitment go hand in hand. The existence of trust for a commitment means that the commitment is not superfluous. The existence of a commitment for trust means that the trust is not misplaced. Chopra and colleagues [14] investigate the connection of trust with architecture. Exploring the above themes further and especially modeling social action as it would arise in future application settings of even moderate complexity would be highly valuable.

6.2 Commitments and Software Engineering

Let me now talk about another important theme with regard to commitments. This has to do with the use of commitments in modeling and realizing multiagent systems in diverse domains. In today's practice, software engineering is mainly concerned with low-level abstractions that are close to implementation details. Such abstractions are difficult to specify and even harder to establish the validity of with respect to the needs of the stakeholders.

Commitments provide a nice alternative basis for specifying software systems. Work on applying commitments for software engineering has been going on for years, since the earliest studies, and initially under the rubric of commitment protocols. However, the more basic challenges of software engineering when applied to interactions in multiagent systems are now beginning to be understood and formulated in terms of commitments [9, 12, 13, 31, 47].

Although the above approaches are useful and promising, they are far from adequate when it comes to the challenges of building systems of practical complexity. I foresee the enhancement of the techniques in terms of clearer specification languages based on commitments, more extensive middleware that supports implementation using abstractions similar to commitments, and the development of tools and technologies to validate and verify commitment-based designs.

In this light, I further think than commitments can inform an expanded notion of norms. Unlike a lot of traditional work, wherein norms are treated as amorphous descriptions of good or normative behavior, I propose that we study norms that like commitments are directed, conditional, contextual, and manipulable. Such norms can help precisely capture normative conditions in a manner where it is clear who is responsible for their enforcement. The notion of organizational context provides a basis for understanding the *governance* of systems of autonomous parties [45], such as service engagements [50] and virtual organizations [2, 48, 49].

7 Conclusions

I have taken this essay as an opportunity to lay out the main themes relating to commitments. I imagine that Professor Castelfranchi and I largely agree with each other on virtually all of the substantial themes regarding commitments. I have highlighted some controversial points in the hope that they would be interesting and useful , especially for those new to the field.

However, to summarize quickly, our points of agreement include the fundamental importance of understanding interaction in multiagent systems from the social and institutional level as opposed to exclusively from the mechanical or operational levels; the very conception of commitments as an elementary social (as opposed to an exclusively mental relationship, as in AI); the distinctions and similarities between practical and dialectical commitments; the value of commitments in understanding institutions and norms; the close relationship between commitments on the one hand and dependence and trust on the other.

Although the field of multiagent systems has made substantial progress since its founding just decades ago, a lot of crucial theoretical and practical problems remain unanswered and even unformulated. No one can predict with any certainty where the field will grow. However, the emergence of networked computing and its expansion into human business and social life suggests that the future of multiagent systems–viewed as the academic field that studies the interactions of social beings–is secure. That our field is now established and has acquired a healthy respect for, if not yet universally a deep understanding of, the social basis for interaction is due in no small part to the imagination and intellect of one researcher and for these invaluable contributions I applaud Cristiano Castelfranchi.

Acknowledgements I have benefited a lot over the years from discussions regarding commitments with a number of people, among them Cristiano himself and, alphabetically, Matthew Arrott, Alexander Artikis, Matteo Baldoni, Cristina Baroglio, Amit Chopra, Marco Colombetti, Nirmit Desai, Frank Dignum, Virginia Dignum, Rino Falcone, Nicoletta Fornara, Les Gasser, Scott Gerard, Paolo Giorgini, Kohei Honda, Michael Huhns, Andrew Jones, Mike Luck, Ashok Mallya, the late Abe Mamdani, Elisa Marengo, Simon Miles, John Mylopoulos, Viviana Patti, Jeremy Pitt, Pankaj Telang, Paolo Torroni, Yathi Udupi, Feng Wan, Michael Winikoff, Jie Xing, Pınar Yolum, and Neil Yorke-Smith. Comments from Michael Huhns, the Dignums, Scott Gerard, Pınar Yolum, and the anonymous reviewer have helped improve this article. I wouldn't presume, however, that any of the people named above agrees with anything I have claimed in this article. I would also like to thank the National Science Foundation for partial support under grant 0910868.

References

1. Artikis A, Sergot MJ, Pitt JV (2009) Specifying norm-governed computational societies. ACM Transactions on Computational Logic 10(1)
2. Brazier F, Dignum F, Dignum V, Huhns MN, Lessner T, Padget J, Quillinan T, Singh MP (2010) Governance of services: A natural function for agents. In: Proceedings of the 8th AAMAS Workshop on Service-Oriented Computing: Agents, Semantics, and Engineering (SOCASE), pp 8–22
3. Castelfranchi C (1993) Commitments: From individual intentions to groups and organizations. In: Proceedings of the AAAI Workshop on AI and Theories of Groups and Organizations: Conceptual and Empirical Research
4. Castelfranchi C (1995) Commitments: From individual intentions to groups and organizations. In: Proceedings of the International Conference on Multiagent Systems, pp 41–48
5. Castelfranchi C (1998) Modelling social action for AI agents. Artificial Intelligence 103(1–2):157–182
6. Castelfranchi C, Falcone R (2010) Trust Theory: A Socio-Cognitive and Computational Model. Agent Technology, John Wiley & Sons, Chichester, UK
7. Castelfranchi C, Falcone R, Marzo F (2006) Being trusted in a social network: Trust as relational capital. In: Trust Management: Proceedings of the iTrust Workshop, Springer, Berlin, LNCS, vol 3986, pp 19–32
8. Chandy KM, Misra J (1986) How processes learn. Distributed Computing 1(1):40–52
9. Cheong C, Winikoff MP (2009) Hermes: Designing flexible and robust agent interactions. In: Dignum V (ed) Handbook of Research on Multi-Agent Systems: Semantics and Dynamics of Organizational Models, IGI Global, Hershey, PA, chap 5, pp 105–139
10. Chopra A, Singh MP (2004) Nonmonotonic commitment machines. In: Dignum F (ed) Advances in Agent Communication: Proceedings of the 2003 AAMAS Workshop on Agent Communication Languages, Springer, LNAI, vol 2922, pp 183–200
11. Chopra AK (2008) Commitment alignment: Semantics, patterns, and decision procedures for distributed computing. PhD thesis, Department of Computer Science, North Carolina State University
12. Chopra AK, Singh MP (2011) Specifying and applying commitment-based business patterns. In: Proceedings of the 10th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), IFAAMAS, Taipei, pp 475–482
13. Chopra AK, Dalpiaz F, Giorgini P, Mylopoulos J (2010) Modeling and reasoning about service-oriented applications via goals and commitments. In: Proceedings of the 22nd International Conference on Advanced Information Systems Engineering (CAiSE), pp 417–421

14. Chopra AK, Paja E, Giorgini P (2011) Sociotechnical trust: An architectural approach. In: Proceedings of the 30th International Conference on Conceptual Modeling (ER), Springer, Brussels, LNCS, vol 6998, pp 104–117
15. Chopra AK, Artikis A, Bentahar J, Colombetti M, Dignum F, Fornara N, Jones AJI, Singh MP, Yolum P (2013) Research directions in agent communication. ACM Transactions on Intelligent Systems and Technology (TIST) In press
16. Desai N, Narendra NC, Singh MP (2008) Checking correctness of business contracts via commitments. In: Proceedings of the 7th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), IFAAMAS, Estoril, Portugal, pp 787–794
17. Falcone R, Castelfranchi C (2009) From dependence networks to trust networks. In: Proceedings of the 11th AAMAS Workshop on Trust in Agent Societies (Trust), pp 13–26
18. Falcone R, Castelfranchi C (2010) Trust and transitivity: A complex deceptive relationship. In: Proceedings of the 12th AAMAS Workshop on Trust in Agent Societies (Trust), pp 43–54
19. Fornara N, Colombetti M (2002) Operational specification of a commitment-based agent communication language. In: Proceedings of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), ACM Press, Melbourne, pp 535–542
20. Fornara N, Colombetti M (2003) Defining interaction protocols using a commitment-based agent communication language. In: Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), ACM Press, Melbourne, pp 520–527
21. Fornara N, Colombetti M (2009) Specifying and enforcing norms in artificial institutions. In: Declarative Agent Languages and Technologies VI, Revised Selected and Invited Papers, Springer, Berlin, LNCS, vol 5397, pp 1–17
22. Gasser L (1991) Social conceptions of knowledge and action: DAI foundations and open systems semantics. Artificial Intelligence 47(1–3):107–138
23. Grosz B, Kraus S (1993) Collaborative plans for group activities. In: Proceedings of the Twelfth International Joint Conference on Artificial Intelligence, pp 367–373
24. Habermas J (1984) The Theory of Communicative Action, volumes 1 and 2. Polity Press, Cambridge, UK
25. Halpern JY, Moses YO (1990) Knowledge and common knowledge in a distributed environment. Journal of the Association for Computing Machinery 37:549–587
26. Herrestad H, Krogh C (1995) Obligations directed from bearers to counterparties. In: Proceedings of the 5th International Conference on Artificial Intelligence and Law, pp 210–218
27. Johnson M, Bradshaw JM, Feltovich PJ, Jonker CM, van Riemsdijk MB, Sierhuis M (2010) The fundamental principle of coactive design: Interdependence must shape autonomy. In: Proceedings of the AAMAS Workshop on Coordination, Organization, Institutions and Norms (COIN), Springer, Toronto, LNCS, vol 6541, pp 172–191
28. Jones AJI, Parent X (2007) A convention-based approach to agent communication languages. Group Decision and Negotiation 16(2):101–141
29. Jones AJI, Sergot MJ (1993) On the characterisation of law and computer systems: the normative systems perspective. In Deontic Logic in Computer Science: Normative System Specification. J. Wiley and Sons, 275–307
30. Levesque HJ, Cohen PR, Nunes JT (1990) On acting together. In: Proceedings of the National Conference on Artificial Intelligence, pp 94–99
31. Marengo E, Baldoni M, Chopra AK, Baroglio C, Patti V, Singh MP (2011) Commitments with regulations: Reasoning about safety and control in Regula. In: Proceedings of the 10th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), IFAAMAS, Taipei, pp 467–474
32. McBurney P, Parsons S (2003) Dialogue game protocols. In: Huget MP (ed) Communication in Multiagent Systems: Agent Communication Languages and Conversation Policies, LNAI, vol 2650, Springer, Berlin, pp 269–283

33. Minsky NH, Ungureanu V (2000) Law-governed interaction: A coordination and control mechanism for heterogeneous distributed systems. ACM Transactions on Software Engineering and Methodology (TOSEM) 9(3):273–305
34. Quine WvO (1960) Word and Object. MIT Press, Cambridge, MA
35. Sacerdoti E (1977) The Structure of Plans and Behavior. Elsevier North-Holland, New York
36. Sichman JS, Conte R, Demazeau Y, Castelfranchi C (1994) A social reasoning mechanism based on dependence networks. In: Proceedings of the 11th European Conference on Artificial Intelligence, pp 188–192
37. Simon HA (1997) Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations, 4th edn. Free Press, New York
38. Singh MP (1991) Social and psychological commitments in multiagent systems. In: AAAI Fall Symposium on Knowledge and Action at Social and Organizational Levels, pp 104–106
39. Singh MP (1998) Agent communication languages: Rethinking the principles. IEEE Computer 31(12):40–47
40. Singh MP (1999) An ontology for commitments in multiagent systems: Toward a unification of normative concepts. Artificial Intelligence and Law 7(1):97–113
41. Singh MP (2000) A social semantics for agent communication languages. In: Proceedings of the 1999 IJCAI Workshop on Agent Communication Languages, Springer, Berlin, Lecture Notes in Artificial Intelligence, vol 1916, pp 31–45
42. Singh MP (2007) Formalizing communication protocols for multiagent systems. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), IJCAI, Hyderabad, pp 1519–1524
43. Singh MP (2008) Semantical considerations on dialectical and practical commitments. In: Proceedings of the 23rd Conference on Artificial Intelligence (AAAI), AAAI Press, Chicago, pp 176–181
44. Singh MP (2011) Trust as dependence: A logical approach. In: Proceedings of the 10th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), IFAAMAS, Taipei, pp 863–870
45. Singh MP (2014) Norms as a basis for governing sociotechnical systems. ACM Transactions on Intelligent Systems and Technology (TIST), In press
46. Singh MP, Chopra AK, Desai N (2009) Commitment-based service-oriented architecture. IEEE Computer 42(11):72–79
47. Telang PR, Singh MP (2011) Specifying and verifying cross-organizational business models: An agent-oriented approach. IEEE Transactions on Services Computing 4, in press
48. Udupi YB, Singh MP (2006) Contract enactment in virtual organizations: A commitment-based approach. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI), AAAI Press, Boston, pp 722–727
49. Udupi YB, Singh MP (2006) Multiagent policy architecture for virtual business organizations. In: Proceedings of the 3rd IEEE International Conference on Services Computing (SCC), IEEE Computer Society, Chicago, pp 44–51
50. Udupi YB, Singh MP (2007) Governance of cross-organizational service agreements: A policy-based approach. In: Proceedings of the 4th IEEE International Conference on Services Computing (SCC), IEEE Computer Society, Salt Lake City, pp 36–43
51. Winikoff M, Liu W, Harland J (2005) Enhancing commitment machines. In: Proceedings of the 2nd International Workshop on Declarative Agent Languages and Technologies (DALT), Springer, Berlin, LNAI, vol 3476, pp 198–220
52. Xing J, Singh MP (2003) Engineering commitment-based multiagent systems: A temporal logic approach. In: Proceedings of the 2nd International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS), ACM Press, Melbourne, pp 891–898
53. Yolum P, Singh MP (2002) Commitment machines. In: Proceedings of the 8th International Workshop on Agent Theories, Architectures, and Languages (ATAL 2001), Springer, Seattle, LNAI, vol 2333, pp 235–247