A formal framework for accountable agent interactions

Carles Sierra, Pablo Noriega
IIIA–CSIC
Campus UAB–Bellaterra, 08193 Barcelona, Spain.
{sierra,pablo}@iiia.csic.es

Abstract

Agent based technologies may constitute a crucial technology for electronic commerce, but only as long as they can inspire adequate confidence to their potential users. We propose to contribute in this direction by developing a notion of "accountability" of agent interactions, which we base on two main constitutive elements: agent-mediated institutions, and shielded agents. In this paper we motivate the need for this type of accountability, and discuss examples of typical features that such accountable interactions should exhibit. We introduce a formal framework to define agent-mediated institutions and develop means for shielding agents in order to characterize accountable interactions and then show how relevant features can be identified and tested. We illustrate the applicability of these ideas in an agent-mediated electronic market place based on the traditional fish market.

Intuitively, an agent-mediated institution is the computational realisation of a set of explicit enforceable restrictions imposed on a collection of dialogical agent types that concur in space and time to perform a finite repertoire of satisfiable actions.

For that characterisation, we will assume that agent interactions are systematically linked to illocutions that are comprehensible to participants and refer to a basic shared ontology. We furthermore assume that all agent interactions are performed according to a pre-established *interaction protocol*, and that other explicit constraints on illocutions and their intended usage and signification can be coded into the *rules of behaviour* to which participating agents are to sumbit. Consequently,

Definition 1 *An **Institution**, I, is a 3-tuple $I = \langle \mathcal{DF}, \mathcal{PS}, \mathcal{BR} \rangle$, where,*

1. $\mathcal{DF}$ *is a* dialogical framework
2. $\mathcal{PS}$ *is a* performative structure,
3. $\mathcal{BR}$ *are the* rules of behaviour *to which participating agents are subject to.*

The *Dialogical Framework*, $\mathcal{DF} = \langle \mathbf{Agents}, \mathbf{Roles}, \mathbf{SocialStructure}, \mathcal{CL}, \mathcal{L}, \mathbf{Time} \rangle$ captures the intuition of *context*. It makes explicit, on one hand, the participants and their basic roles, as well as their relevant social interrelationships. On the other, it also makes explicit the communication and object languages, $\mathcal{CL}$, $\mathcal{L}$ that will be neded for illocutions to be shared between particpating agents, as well as a common notion of time to which sequencing of interactions may need to refer. Note, however, that nothing is said about the internal components of participating agents in this framework, only general rules of behaviour are later on prescribed.

A *performative structure*, $\mathcal{PS} = \langle \mathbf{S}, \mathbf{SDG} \rangle$, is a set of interdependent scenes. Each scene is defined as a set of agents who are each to assume a given *role*, all of the agents are subject to a common *interaction protocol*. Protocols are finite state machines where state transitions are labelled by illocutions and states have associated memory stacks (of "commitments"). We use a *Scene Dependence Graph* to establish causal and temporal co-dependencies among initial and terminal state commitments of different scenes.

$\mathcal{BR}$ captures the notion of *intended behaviour* by assigning to each participating agent a collection of schemata of the form:

$$\mathbf{IF}(\psi_1 \wedge \ldots \wedge \psi_n)\ \mathbf{THEN}\ \xi$$

where $\xi \in \mathcal{CL}$, and $\psi_1, \ldots \psi_n \in \mathcal{L} \cup \mathcal{CL}$. These schemata should ideally be part of the agent's internal theory, in order to guarantee that the agent upholds the institutional conventions. To get a handle on this issue, we introduce the notion of *governor* —or *co-agent*— to denote an agent-like entity c_a that enforces the rules of behaviour for an agent-type role onto a specific agent a.

Given an agent a of role $Role(a)$, whose rules of behaviour are $\mathcal{BR}(Role(a))$, c_a actual behaviour should be consitent with those rules in the sense that for whatever illocution $\iota(a, b, \phi, t)$, if it is *required* by the behaviour rules, it will be uttered by the agent–co-agent pair, and if $\iota(a, b, \phi, t)$ is uttered by the agent–co-agent pair, it is not forbidden by the rules of behaviour. These consistency conventions may be stated as follows:

Definition 2 *Given an agent a of type $Role(a)$ and subject to a a set of rules of behaviour $\mathcal{BR}(Role(a))$,* a co-agent c_a *for a, is the minimal agent such that c_a upholds $\mathcal{BR}(Role(a))$ for a. I.e., if for every illocution $\iota(a, b, \phi, t)$*

1. $\mathcal{BR}(Role(a)) \vdash \Box\iota(a, b, \phi, t) \Longrightarrow T_{a \cup c_a} \vdash \iota(a, b, \phi, t)$, *and*
2. $T_{a \cup c_a} \vdash \iota(a, b, \phi, t) \Longrightarrow B_{rules}(Role(a)) \not\vdash \Diamond\neg\iota(a, b, \phi, t)$

And then properly implemented. Intuitivelly, we will have a pair of agents —an agent a and its co-agent c_a — acting as one. The co-agent c_a filters all incoming and outgoing illocutions, and in general guarantees that all rules of behaviour associated with the agent's type are actually met. Thus, c_a receives all incoming illocutions and re-sends them to a who may deliberate on them. In the meanwhile, the co-agent deliberates and prepares whatever illocutionary actions may be required by the rules of behaviour, and also identifies those that may be consistent with it. When, after its deliberation, a utters an illocution, it is filtered by c_a –*i.e,* if the illocution is appropriate it is re-uttered by the co-agent, and if the illocution is inappropriate the co-agent inhibits it, and gives the agent an indication of failure– and the pair c_a–a proceeds to a new state. Hhowever, if a triggering condition is met —*e.g.*, if there is a time constraint for a response— and a has not been able to produce a *required answer*, then c_a should provide a default answer to guarantee compliance with the existing protocol and rules of behaviour and informs the agent of its execution.