Institutions in Channel Theory

Stijn De Saeger
Japan Advanced Institute of Science and Technology
stijn@jaist.ac.jp

July 31, 2007

Abstract

In this talk, we propose an account of institutions based on channel theory, and show that the key elements in the theory (classifications, local logics) offer an expressive framework for the formalization of institutional action.

1 Introduction

Multiagent systems have been characterized as *technological extensions of human society* in [4]. To the extent that many of the problems encountered in formalizing the behaviour of software agents do indeed have close analogues in the "real" world, many aspects of agent societies have been modeled after their real world counterparts, drawing on a wealth of available research in such fields as epistemic logic, game theory, belief revision, and so on.

In order to formalize communication between (artificial) agents, a number of different agent communication languages (ACLs) have been proposed in the artificial intelligence community. Despite individual differences, they mostly seem to agree that an appropriate formalization of the theory of *speech acts*, pioneered by Searle [5] and Austin [1], is the way to go. This is hardly surprising: the concept of speech acts has a solid philosophical foundation, and it allows to cash out an account of agent communication in terms of *rational action*, making it an extension of an already well-studied problem in AI.

When it comes to giving a semantics for these ACLs, the dominant approach has been a *mentalistic* or *intention-based* one, in the tradition of the Gricean *intention recognition* model of cognition (see work by Cohen & Levesque, Bratman). A prime example is the definition of "inform" speech acts in terms of the feasibility preconditions (**FP**) and postconditions (**RE** or "rational effect") on agents' belief states in the FIPA specification of agent communication. [1]

[1] http://www.fipa.org/repository/aclspecs.html

$$
\begin{array}{ll}
< \mathsf{inform}(i, j, \varphi) > & \\
\quad [\mathbf{FP}] & B_i\varphi \wedge \neg B_i(B_j\varphi \vee B_j\neg\varphi) \\
\quad [\mathbf{RE}] & B_j\varphi
\end{array}
\tag{1}
$$

The preconditions say that i should believe φ, and should also believe that j does not already have an opinion on whether or not φ. In those situations, the effect of an inform act is that j comes to believe φ as well. As some researchers have claimed, this approach faces some issues. Besides long-standing problems with the formalization of intensional concepts like belief (e.g. logical omniscience), there is a tension between the essentially public nature of communication and the private nature of agent beliefs. Communication is by its nature a social activity, and defining its meaning in terms of subjective mental states that are in principle not open for inspection (either by other agents or by some system-level observer) seems to miss an important conceptual point: Belief updates fail to capture the *social* updates triggered by speech acts. If i's inform action goes uncontested by j, i is from that point on entitled to treat j "as if" she believed φ, according to the implicit rules of the dialogue game.

In response to the problems faced by mentalistic theories of agent communication, a number of researchers (Colombetti, [3] and Singh, [6]) have advocated to ground a theory of speech acts in the *social* updates they effect, rather than the epistemic ones — essentially defining an act by agent i of informing agent j that φ as committing i to the truth of φ, vis-a-vis j. Such a *social* semantics for ACLs is appealing for a number of reasons, essentially replacing the thorny issue of talking about the private beliefs of agents by the more transparent notion of adherence to social norms and commitments [2]. At the same time though, it places speech act semantics firmly in the realm of *institutions*:

> [...] "institutions" are systems of constitutive rules. Every institutional fact is underlain by a (system of) rule(s) of the form "X counts as Y in context C". (J. Searle, [5] p.51)

As argued convincingly in [3], institutional action and natural action have profound differences, which poses a new set of problems for traditional AI theories of rational action. AI theories like BDI logics are traditionally defined in terms of agent intentions and processes of physical causation between events: agent i forms an intention to update agent j's belief state with φ, and this intention itself sets in motion a causal chain of events leading to its fulfillment. However, the success (or otherwise) of an institutional action depends on parameters beyond the control of the agents involved — institutions crucially involve some kind of *collective intensionality* that transcends any single instance of such event. For instance, asking for a beer counts as a commitment to purchase in

[2] Not to be confused with the mentalistic notion of commitment as a kind of "persistent intention".

the context of pubs because the institute of commerce *says* that it does. As Searle argues, these bodies of social convention are not *normative* rules, which are concerned with the distribution of obligations and permissions over a society of agents. Instead they are in a very real sense *constitutive* of what we recognize as successful social action, through the distribution of institutional powers among agents. In order to develop a proper social semantics of speech acts, we need a formal account of institutions and their logical properties.

2 An Account of Institutions

How are we to make sense formally of a statement like "X counts as Y in context C" (and the ways it might fail to hold). As a first approximation, we say a given event e supports an institutional fact Y in a context C when:

i. e has a physical property X, such that

ii. X is a proxy for Y by virtue of some institutional context C, in which

iii. "$X \Rightarrow_c Y$" is a constitutive rule of C.

Channel Theory We propose to use *channel theory* (Barwise & Seligman, [2]) as a platform for formalizing institutional action. The main objects in channel theory are domain classifications, morphisms between these classifications, and the logics they give rise to. Let $\mathcal{C}_P$ be a classification representing *physical* reality. Formally, $\mathcal{C}_P$ is a triple $\langle S_P, \Sigma_P, \models_P \rangle$ classifying so-called "brute facts" in S_P (i.e. a set of situation or event *tokens* having a spatio-temporal extension[3]) according to natural event *types* in Σ_P, which is expressed in the binary classification relation $\models_P \subseteq S_P \times \Sigma_P$. Thus for $s \in S_P$, $raiseHand_i \in \Sigma_P$, we write

$$s \models_P raiseHand_i \tag{2}$$

if s is an event of agent i raising her hand.

On the other hand there is the realm of *social* reality, which we may represent similarly using a classification $\mathcal{C}_S = \langle S_S, \Sigma_S, \models_S \rangle$, classifying situations S_S according to their social meanings Σ_S in a relation $\models_S \subseteq S_S \times \Sigma_S$. For instance, the following instance of this relation says that s' is a situation in which agent i makes a bid.

$$s' \models_S makesBid_i \tag{3}$$

Institutions are systems of constitutive rules, i.e. constraints linking physical actions to their social implications. They account for the fact that the event s of i raising her hand in $\mathcal{C}_P$ may count as an event s' of making a bid in the context of auctions, but also as an event of, say, volunteering to solve a problem in math class.

[3] For our purposes, nothing substantial hangs on the distinction between *situations* and *events*, so we treat them as one.

$$\begin{tikzcd}
& \Sigma_P \cup \Sigma_S \arrow[d, dash, "\models_I"] & \\
\Sigma_P \arrow[ur, "f^\wedge"] \arrow[d, dash, "\models_P"] & S_P \times S_S \arrow[dl, "f^\vee"'] \arrow[dr, "g^\vee"] & \Sigma_S \arrow[ul, "g^\wedge"'] \arrow[d, dash, "\models_S"] \\
S_P & & S_S
\end{tikzcd}$$

Figure 1: Institutions as theories on $\mathcal{C}_I$

Classifications give rise to a hierarchy of contexts or *local logics*. Let a pair of sets $\Gamma, \Delta \subseteq \Sigma_P$ be a *constraint* on a situation c (written "$\Gamma \vdash_c \Delta$") iff when $c \models X$ for all $X \in \Gamma$ then $c \models Y$ for some $Y \in \Delta$. By extension, $\Gamma \vdash_C \Delta$ iff $\Gamma \vdash_c \Delta$ for all $c \in C \subseteq S$. For example, "$scratchHead_i \vdash_{S_P} raiseHand_i$" is a constraint holding for all situations in S_P. Similarly, constraints on the social classification $\mathcal{C}_S$ (like for instance "$makeBid_i \vdash_S mustPurchase_i$") represent normative rules. A context $\langle \mathcal{C}, \vdash, N \rangle$ on a classification $\mathcal{C}$ then consists of a set of constraints $\vdash$ (closed under logical consequence) and a set of situations $N \subseteq S$ that are considered "normal" for this context, meaning that they satisfy all constraints in $\vdash$.

Institutions then are theory-like objects stipulating how a piece of information in $\mathcal{C}_P$ relates to $\mathcal{C}_S$ in a principled way. This suggests a *binary channel* construction [4] (Fig. 1) around an institutional classification $\mathcal{C}_I = \langle S_I, \Sigma_I, \models_I \rangle$, where S_I is the Cartesian product of S_P and S_S, i.e. a set of pairs $\langle s_1, s_2 \rangle$ (for $s_1 \in S_P, s_2 \in S_S$ [5]) and Σ_I is the disjoint union of Σ_P and Σ_S. Let $\langle f^\vee, f^\wedge \rangle$ and $\langle g^\vee, g^\wedge \rangle$ be the usual *infomorphisms* [6], such that $\langle s_0, s_1 \rangle \models_I X$ iff $X = f^\wedge(X')$ and $s_0 \models_P X'$ (resp. $X = g^\wedge(X')$ and $s_1 \models_S X'$). We call a set of constraints $\{\langle \Gamma, \Delta \rangle \mid \Gamma, \Delta \subseteq \Sigma_I\}$ on $\mathcal{C}_I$ closed under identity, weakening and cut a body of constitutive rules, and a context $\langle \mathcal{C}_I, \vdash_I, N_I \rangle$ on $\mathcal{C}_I$ an *institution*. Then institutions act as *theories* on the alignment of $\mathcal{C}_p$ and $\mathcal{C}_S$.

We can now give formal substance to the claim that i's raising a hand counts as making a bid in the social context $A \in \mathsf{CXT}(\mathcal{C}_S)$ of auctions, namely as a constraint of the corresponding institution:

$$f^\wedge(raiseHand_i) \vdash_{g[A]} g^\wedge(makeBid_i)$$

In this picture, any instance s of a "hand raising" event counts as an act of "bidding" in a social context A if every interpretation $\langle s, s' \rangle$ of s in a social situation s' of context A is normal with respect to the constraint $f^\wedge(raiseHand_i) \vdash_{g[A]} g^\wedge(makeBid_i)$.

[4] See [2].

[5] Note that a pair $\langle s_1, s_2 \rangle$ does not *necessarily* denote two distinct situations. They could be different perspectives (i.e. "physical" and "social") on the same event.

[6] Given classifications $\mathcal{C}_A$ and $\mathcal{C}_B$, an infomorphism $f : \mathcal{C}_A \rightleftarrows \mathcal{C}_B$ from $\mathcal{C}_A$ to $\mathcal{C}_B$ is a pair of contravariant functions $\langle f^\wedge, f^\vee \rangle$ satisfying that $\forall s \in S_B,\ \sigma \in \Sigma_A : f^\vee(s) \models_A \sigma$ iff $s \models_B f^\wedge(\sigma)$. (Again see [2])

Count-as Conditionals and Nonmonotonicity Modeling count-as conditionals as the constraints of some local logic allows us to account for the essential context-dependence of constitutive norms. For example, $f^\wedge(raiseHand_i) \vdash_{g[A]} g^\wedge(makeBid_i)$ may be a valid constitutive rule in the institutional context $g[A]$ of auctions, while still allowing the raising of a hand to count as some other act in another social context (say, "$f^\wedge(raiseHand_i) \vdash_{g[V]} g^\wedge(vote_i)$" in the context V of voting). One well-known logical property of count-as conditionals is their nonmonotonicity. Constitutive rules like $raiseHand_i \vdash makeBid_i$ generally do not admit left strengthening (eg. $raiseHand_i, scrathHead_i \nvdash makeBid_i$) or right weakening (eg. $raiseHand_i \nvdash makeBid_i, disownsChildren_i$).

In channel theory, nonmonotonicity is dealt with at the level of contexts, rather than rules. That is, *inside* a given context all inferences are monotonic, by default. Those instances of left strengthening or right weakening that are problematic are cases where a strengthened (respectively weakened) constraint turns out to be in conflict with some situations $N' \subseteq N$ assumed normal with respect to this context, and would thus warrant a shift to some stronger (weaker) institutional context. [7]

References

[1] J. L. Austin. *How to Do Things With Words.* Oxford University Press, Oxford, 1962.

[2] J. Barwise and J. Seligman. *Information Flow. The Logic of Distributed Systems.* Cambridge Tracts in Theoretical Computer Science, Cambridge University Press, 1997.

[3] M. Colombetti and M. Verdicchio. An analysis of agent speech acts as institutional actions. In *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, pages 1157–1164, New York, NY, USA, 2002. ACM Press.

[4] N. Fornara, F. Viganò, and M. Colombetti. Agent communication and institutional reality. In *International Workshop on Agent Communication AC2004*, pages 1–17, 2004.

[5] J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press, Cambridge, 1969.

[6] M. P. Singh. Agent communication languages: Rethinking the principles. *Computer*, 31(12):40–47, 1998.

[7] As shown in [2], the set of contexts of a classification $\mathcal{C}$ forms a complete lattice under context subsumption ($\sqsubseteq$), where $\langle \mathcal{C}, \vdash_1, N_1 \rangle \sqsubseteq \langle \mathcal{C}, \vdash_2, N_2 \rangle$ iff $\vdash_1 \subseteq \vdash_2$ and $N_1 \supseteq N_2$. A context strengthening or weakening then corresponds to moving to some weaker or stronger context in the $\sqsubseteq$ hierarchy.