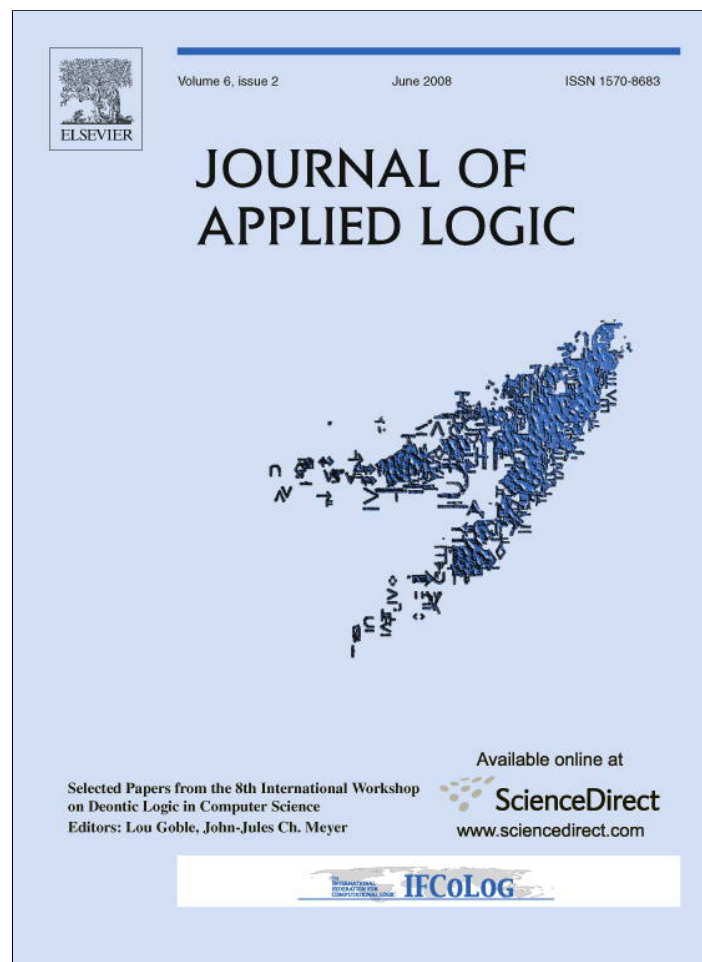


Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



# The normative aspect of signalling and the distinction between performative and constative

Andrew J.I. Jones<sup>a,\*</sup>, Steven O. Kimbrough<sup>b</sup>

<sup>a</sup> King's College London, London, UK,  
Department of Computer Science

<sup>b</sup> Operations & Information Management,  
University of Pennsylvania, Philadelphia, USA

Available online 29 June 2007

---

## Abstract

The paper outlines an approach to the formal representation of signalling conventions, emphasising the prominent role played therein by a particular type of normative modality. It is then argued that, in terms of inferencing related to this modality, a solution can be given to the task J.L. Austin set but failed to resolve: finding a criterion for distinguishing between what Austin called *constatives* and *performatives*. The remainder of the paper indicates the importance of the normative modality in understanding a closely related issue: reasoning about trust in communication scenarios; this, in turn, facilitates a clear formal articulation of the role of a *Trusted Third Party* in trade communication.

© 2007 Elsevier B.V. All rights reserved.

*Keywords:* Agent communication languages; Speech act theory; Applied modal logic; Trust; Business communication; Trusted-third-party

---

## 1. Introduction

The approach to the analysis of communicative acts taken in this paper differs from those currently most in vogue, in that its focus is neither on the intentions of communicators (FIPA: <http://www.fipa.org/>, and in particular <http://www.fipa.org/repository/bysubject.html> and <http://www.fipa.org/repository/aclspecs.html><sup>1</sup>) nor on their supposed commitments [1,2]. By contrast, the focus here is on the conventions that—as we shall say—*constitute* any given communication system *s*. These conventions make possible the performance of meaningful communicative acts by the agents, human or electronic, who have adopted *s* as a means of communicating with each other. We begin by summarising some of the main features of the approach.

---

\* Corresponding author.

*E-mail addresses:* [andrewji.jones@kcl.ac.uk](mailto:andrewji.jones@kcl.ac.uk) (A.J.I. Jones), [kimbrough@wharton.upenn.edu](mailto:kimbrough@wharton.upenn.edu) (S.O. Kimbrough).

<sup>1</sup> Accessed 2007-05-27.

## 2. Signalling conventions

A convention-based system that defines a framework for agent interaction may appropriately be called an *institution*.<sup>2</sup> In common with other institutions, communication systems exist to serve a purpose; specifically, their purpose, or point, obviously, is to facilitate the transmission of information of various kinds.

In order to develop these intuitions, and to begin to move towards a formal model, we look first at the communicative act of *asserting* (or *stating*, or *saying*) *that such-and-such is the case*. The key question is this: in the constitution of communication system/institution  $s$ , what is it that makes it possible for an agent, if he so wishes, to make an assertion? Our answer is that  $s$  contains conventions according to which the performance of particular acts *count as* assertions, and which also specify what those acts mean. Consider, by way of illustration, the institution that was once operative for sea-going vessels, in virtue of which they were able to send signals indicating aspects of the state of a vessel by hoisting sequences of flags. Raising flag sequence  $q1$  would count (by convention) as a means of saying that the vessel was carrying explosives, raising flag-sequence  $q2$  would conventionally count as indicating that the vessel carried injured crew members. . . and so on. Note the general form of the conventions themselves: they each associate a particular type of act with a particular state of affairs, and because they are conventions for *asserting* (i.e., for *that* type of communicative act) they each count as a means of *saying that* the associated state of affairs holds.

For present purposes, it matters not at all which sorts of acts are used in a given communication system; the account of communication conventions we offer is entirely neutral on that issue.<sup>3</sup>

Suppose now that in communication system/institution  $s$ , the act of bringing it about that  $A$  counts as a means of asserting that the state of affairs described by  $B$  obtains (abbreviating: by convention in  $s$ , doing  $A$  counts as an assertion that  $B$ ). And suppose further that agent  $j$ , who is an  $s$ -user, does  $A$  in circumstances in which  $B$  does not hold.<sup>4</sup> Then it is appropriate to say that, from the point of view of the institution  $s$ , something has gone wrong, *in as much as the purpose or function within institution  $s$  of acts of asserting is to facilitate the transmission of reliable information*. The point of asserting, as an institutionalised act, is to be able to show how things stand in a given state of affairs. Given that this *is* the point of asserting, the doing of  $A$  in circumstances where  $B$  does not hold is a form of *abuse* of the system. Relative to the purpose of asserting, as an institutionalised act,  $A$  *ought* to be done *only* when  $B$  is the case, and so the doing of  $A$  in non- $B$  circumstances amounts to a deviation from the ideal that the system is supposed to achieve.<sup>5</sup>

The conventions for asserting make it possible for acts of assertion to be performed, and they do so by indicating what *would* be the case in circumstances in which the purpose of asserting, qua institutionalised act, is fulfilled. If, by convention in  $s$ , doing  $A$  counts as an assertion that  $B$ , then in ideal circumstances (with respect to  $s$ )  $B$  holds whenever  $A$  is done. These observations are the key to understanding the intuitions on which is grounded the general logical form we assign to communication conventions of the assertoric type.<sup>6</sup>

Following the theory developed in [3] and [4], the form of the signalling convention (sc) according to which, in  $s$ , agent  $j$ 's seeing to it that  $A$  counts as an assertion that  $B$ , is given by

$$(\text{sc-assert}) \quad E_j A \Rightarrow_s I_s^* B$$

where expressions of the form  $E_j A$  are read ' $j$  sees to it that  $A$ ',  $\Rightarrow_s$  is the 'counts as' connective of [9], and  $I_s^*$  is a normative operator, intended to capture the sense of 'ought', or ideality, alluded to above. Details of the logics

<sup>2</sup> This section offers merely a summary of the approach to signalling conventions described in [3] and [4], and the reader is referred to those sources—particularly to [4]—for a more detailed account, including details of the component modalities. The focus of the present paper is on the *normative* aspect of signalling conventions, and its relation to *trust* and to Austin's distinction between *performative* and *constative*.

<sup>3</sup> By 'communication system' we here mean the set of conventions that constitute the system, together with the set of agents who make use of those conventions.

<sup>4</sup> It is irrelevant to the present point whether or not  $j$  believes that  $B$  does not hold.

<sup>5</sup> One of the reviewers helpfully pointed out that this way of characterising the purpose of asserting strongly suggests an analysis of 'ought' along Andersonian lines, relating failure to violation. In fact, in the first-order simplification of our account outlined in [5]—see below Section 5—a reduction of a similar kind is indeed proposed.

<sup>6</sup> [6] is the source from which we take the idea that, in order to understand the communicative act of asserting, one must understand in what sense of 'ought' that which is asserted *ought to be true*. Stenius's much neglected paper is in our opinion one of the most insightful essays written on the analysis of different types of communicative acts. The idea that the 'counts as' notion figures crucially in the convention constituting asserting appears for the first time, to our knowledge, in [7]. For further discussion of the philosophical roots of our approach, see [8].

and semantics for the action and ‘counts as’ modalities are given in [9].<sup>7</sup> Expressions of the form (sc-assert) say that  $j$ ’s seeing to it that  $A$  counts in conventional signalling system  $s$  as a means of indicating that, were  $s$  to be in an ideal/optimal state with respect to its function of facilitating the transmission of reliable information,  $B$  would be true.

The logic of the normative modality is that of a (relativised) normal modality of type K. Closure under logical consequence is a natural assumption, given the intended interpretation of the operator, for if a signalling system would be in an ideal state only if  $B$  were true, then it would be in an ideal state only if the logical consequences of  $B$  were also true. Note also (cf. [10, p.184]) that the absence of the D. schema reflects the obvious fact that mutually inconsistent assertions can be made: according to one of them,  $B$  ought to be true, but according to the other,  $B$  ought to be false.

Such other types of communicative acts as *commanding*, *promising*, *requesting* and *declaring* (the latter in the sense of [11]) are characterised in terms of signalling conventions of the same basic form as that of (sc-assert) with, crucially, some further elaboration of the scope-formula  $B$  falling to the immediate right of the  $I_s^*$  operator in the consequent (cf. [3] and [4]). This means, of course, that each of these communicative act-types is here treated as a sub-species of the act of asserting, a consequence of the fact that—in stark contrast to Austin [12]—we take all communicative acts to be acts of transmitting signals which may, or may not, be true. We shall see in due course how this approach provides the basis for formally articulating the distinction that Austin sought, but failed to capture, between what he called *constatives* and *performatives*.

The form of the signalling convention for *commanding* is

$$\text{(sc-command)} \quad E_j A \Rightarrow_s I_s^* O E_k B$$

where the  $O$  operator is a directive normative modality representing obligation. (We do not here specify a logic of obligation, since it is not the focus of our concern. For present purposes, SDL (standard deontic logic) would suffice.) According to (sc-command), if  $j$  sees to it that  $A$ ,  $s$  would then be in an ideal state (things would then be as they ought to be), relative to  $s$ ’s function of facilitating the transmission of reliable information, if there were then an obligation on  $k$  (the agent to whom the command is addressed) to see to it that  $B$  (where  $B$  is the state of affairs that  $k$  is commanded to bring about).

The form of the signalling convention for *promising* is

$$\text{(sc-promise)} \quad E_j A \Rightarrow_s I_s^* O E_j B$$

According to (sc-promise), if  $j$  sees to it that  $A$ ,  $s$  would then be in an ideal state (things would then be as they ought to be), relative to  $s$ ’s function of facilitating the transmission of reliable information, if there were then an obligation on  $j$  (the agent making the promise) to see to it that  $B$  (where  $B$  is the state of affairs that  $j$  promises to bring about).<sup>8</sup>

The form of the signalling convention for *requesting* is

$$\text{(sc-request)} \quad E_j A \Rightarrow_s I_s^* H_j E_k B$$

where expressions of the form  $H_j A$  are read ‘ $j$  attempts to see to it that  $A$ ’, and the logic of the *attempts* operator is essentially that of the action operator *minus* the ‘success’ condition (the T. schema). According to (sc-request), if  $j$  sees to it that  $A$ ,  $s$  would then be in an ideal state (things would then be as they ought to be), relative to  $s$ ’s function of facilitating the transmission of reliable information, if  $j$  were attempting to get  $k$  to see to it that  $B$ .

The point of declaratives is to create a new state of affairs, as when, for instance, a couple are declared married, or a meeting is declared open. Let  $j$  be the agent issuing the declarative, and let  $B$  describe the state of affairs to be created by the performance of the declarative. Then the form of the governing convention is

$$\text{(sc-declare)} \quad E_j A \Rightarrow_s I_s^* E_j B$$

<sup>7</sup> There is, however, a point of difference between the present treatment of ‘counts as’ and the account given in [9], in as much as the role of the ideality operator in our analysis of signalling conventions obviates the need for the D-operator as that was employed in [9].

<sup>8</sup> We accept that a case can be made for inserting the operator  $E_j$  immediately to the left of the obligation operator in the consequent of (sc-command) and (sc-promise), since it is the agent  $j$  who, by performing the communicative act  $E_j A$ , sees to it that the obligation is created. A move of that sort would then make commanding and promising sub-species of declaring (see below), which is perhaps a very natural way of viewing these matters. A change of this kind could be made without necessitating revision of the main points addressed in this paper.

According to (sc-declare), if  $j$  sees to it that  $A$ ,  $s$  would then be in an ideal state (things would then be as they ought to be), relative to  $s$ 's function of facilitating the transmission of reliable information, if it were then the case that  $j$  has indeed seen to it that  $B$ .

### 3. Distinguishing constatives from performatives

Austin sought a grammatical criterion for distinguishing between constative sentences (characteristically used in communicative acts the point of which is essentially to state or assert that such-and-such is the case) and performative sentences, which are characteristically employed—as he saw it—in doing the *other* kinds of things that one does with words, i.e., other than stating/asserting, such as giving orders, accepting offers, making promises, opening meetings and naming ships. The first seven lectures recorded in the posthumously published *How to Do Things with Words* [12] describe his ultimately unsuccessful attempt to define an appropriate distinguishing criterion—a criterion compatible with the basic assumption he made to the effect that performative sentences, unlike constatives, lack truth values. On our view it was in part that very assumption that prevented him from finding what he sought.<sup>9</sup>

We have characterised four types of performatives (*commanding*, *promising*, *requesting* and *declaring*) in terms of conventions that are all special cases of the convention for asserting, (sc-assert). So we are maintaining that the general form of all of these conventions is expressed by (sc-assert). Suppose now that agents  $j$  and  $k$  are users of communication system  $s$ , and that they are mutually aware of the content of the various instances of (sc-assert), each of which shows what the communicative acts performable in  $s$  mean. ( $j$ 's seeing to it that  $A1$  counts as an assertion that  $B1$ ,  $j$ 's seeing to it that  $A2$  counts as an assertion that  $B2$ ,  $j$ 's seeing to it that  $A3$  counts as a command to do  $B3$ ,  $j$ 's seeing to it that  $A4$  counts as a request to do  $B4$  ... and so on. The particular instances of (sc-assert) are, we may say, the code that constitutes  $s$ .)

In terms of the general form of communicative conventions, as expressed by (sc-assert), we may say that  $k$ , on witnessing  $j$ 's performance of the act  $E_j A$ , forms a belief<sup>10</sup> the content of which is the consequent of (sc-assert):<sup>11</sup>

$$B_k I_s^* B \tag{1}$$

The key question now is this: under what conditions would  $k$ , as a rational agent, be prepared to trust the reliability of  $j$ 's communicative act, and move from the belief expressed by (1) to (2)?

$$B_k B \tag{2}$$

Crucially, the answer to this question will depend on whether  $j$ 's act is a performative or a constative, in Austin's sense. If it is a performative (for instance, one of the four types mentioned above) then  $k$  will be justified in making the inference from (1) to (2) provided merely that  $j$  is relevantly *empowered*—i.e., empowered to give commands, or empowered to make requests, or empowered to make promises, or empowered to make declarations. Consider commanding: if  $j$  is empowered/authorised to give commands, then his performance of the communicative act of commanding will indeed create an obligation on the addressee,  $k$ , to see to it that  $B$ . The scope formula to the right of the  $I_s^*$  operator in the convention is made true by  $j$ 's performance of the communicative act. If he is empowered, then 'saying makes it so'.

The situation with respect to constatives, however, is quite different, for here there is no notion of empowerment or authorisation which would *itself* license the inference of  $B$  from  $I_s^* B$ . The closest one could get to such a notion would arise in cases in which  $j$  is deemed to be an authority on the subject about which he is making an assertion. But even then, his *saying* that  $B$  does not in itself *make it the case* that  $B$ . The signal he transmits is not 'verifiable by its use', but by appeal to the facts on which he is deemed to have expert, or authoritative, knowledge.

Does this analysis do justice to a distinction—considered by Austin to be important—between *fully performative* and *merely descriptive* usage of performative sentences? To explain the question, consider the utterance by the officer-

<sup>9</sup> What follows in due course below has its roots, in part at least, in an old idea. A number of early contributors to the literature on performatives (Lemmon, Åqvist and Lewis among them) suggested that the characteristic feature of performatives, in contrast to constatives, was 'verifiability by use', or the fact that 'saying makes it so'. See [8] for references and discussion.

<sup>10</sup> For present purposes we shall assume that the belief modality is assigned the logic of a relativised normal modality of type KD.

<sup>11</sup> This is the default conclusion  $k$  will draw, on the assumption that  $j$ 's act is a serious communicative act, i.e., a literal implementation of the governing (sc-assert) convention. For more detail on this, see [3].

in-command of the performative sentence ‘I command you to open fire’ in two different contexts: in the first, he is using the utterance itself to give the command (the fully performative usage), but in the second he is giving the command by signing a written order, and uttering the sentence ‘I command you to open fire’ so as to describe what (by signing) he is doing. The answer to the question is surely affirmative, for the difference between the two cases lies precisely in the evidence that would be required in order to justify inferring that an obligation (to open fire) had been created. For the fully performative case, the inference is justified if the communicator is indeed empowered to issue commands. But in the descriptive case more evidence is needed, for there the inference is justified only if it is the case *both* that the communicator is empowered to command *and* that he is performing another action by means of which he is exercising that authority (signing the written order). The descriptive case falls then in the category of constatives, according to our criterion, and this is surely in line with the point Austin had in mind regarding these different usages of performative sentences.

Towards the end of Lecture VII in [12], Austin gives up the pursuit of a distinguishing criterion. He says this:

Now we failed to find a grammatical criterion for performatives, but we thought that perhaps we could insist that every performative *could* be in principle put into the form of an explicit performative, and then we could make a list of performative verbs. Since then we have found, however, that it is often not easy to be sure that, even when it is apparently in explicit form, an utterance is performative or that it is not; and typically anyway, we still have utterances beginning ‘I state that ...’ which seem to satisfy the requirements of being performative, yet which surely are the making of statements, and surely are essentially true or false.

It is time to make a fresh start on the problem. We want to reconsider more generally the senses in which to say something may be to do something, or in saying something we do something (and also perhaps to consider the different case in which *by* saying something we do something). [12, p. 91]

And then it is in the remaining lectures that Austin developed the now familiar distinction between *locutionary*, *illocutionary* and *perlocutionary*—indeed the latter two are already hinted at in the last bit of the passage just quoted. On our view, by contrast, there is no need to despair of finding a means of distinguishing constative from performative, but one should look not for a grammatical criterion, as Austin did, but at the grounds upon which one may justifiably infer a belief of form (2), above, from a belief of form (1).

There is also no need to resort to the theory of illocutionary acts; for we can supply a formal characterisation of different types of communicative acts—as outlined above—that makes no explicit use of the notion of illocutionarity, and which, in contrast to the approach taken by FIPA (see FIPA URLs, cited above), does not focus on the intended perlocutionary effects (what FIPA call the ‘rational effects’) of communication.

As indicated above, we give the analysis in terms of conventions that specify what ought to hold true when, for instance, an order is given or a request or promise is made. The normative, ideality operator is the key element, marking what will be the case if the governing convention is exploited in a way that conforms to the function that the communication/signalling system is designed to fulfil: the transmission of reliable information.

This, in turn, enables us to represent in a very straightforward way the belief of an agent who is aware of what a particular transmitted signal means (see above, formula (1)). The content of that belief is a normative expression, of form  $I_s^*B$ , where  $s$  is the communication system used in transmitting the signal. To be aware of what the signal means, on our view, is just to be aware of what, by convention, ought to be true given that the signal has been sent—it is to be aware of what would be the case if the reliability of the communicator could be trusted. In contrast to some other approaches to the analysis of Agent Communication Languages (ACLs), we do not need to require the recipient to believe that the communicator is intending to produce in him the belief that  $B$  or that the communicator believes that  $B$ , or the belief that the communicator intends to get him to recognise that it is the communicator’s intention to get him to believe that  $B$  ... or indeed any other part of the convoluted Gricean mechanism.<sup>12</sup> Our approach is very much simpler, and is made possible, essentially, by the role played by the normative operator.

We note in passing one additional advantage of our approach. First, it facilitates third party determination of what is said. Conventions, unlike intentions, beliefs, and desires, are quite public and open to objective assessment by disinterested parties. This is a key property if disputes are to be resolved in a manner that discourages cheating and

<sup>12</sup> A considerably more detailed critique of the Gricean approach—in which of course the FIPA approach has its roots—is to be found in [8, Chapter 4].

renegeing. Intentions, beliefs, desires and other mental states are, perhaps, not entirely inaccessible to neutral third parties. Even so, they are quite problematic in comparison to established conventions. This is apparent in the case of commercial transactions and electronic commerce in particular, but the point applies in the large, to all forms of communication for which it is valuable to be able to ascertain what was said in a fair and objective way.

#### 4. Reasoning about messages received

As we have seen, the formal characterisation of the belief state of a message recipient  $k$  enables us to represent what it would be for  $k$  to trust the reliability of the message sent:  $k$  would make the transition from a belief of type (1) to a belief of type (2). The formalism also facilitates the representation of the reasoning of  $k$  in a situation prior to that in which he has decided whether or not to trust messages he has received. This is important at least for the reason that, in trying to determine whether trust is justified,  $k$ —as a rational agent—will want to evaluate the consistency of the messages he has received with other beliefs he already holds.

To illustrate, suppose that  $k$  has received a message asserting that  $B$ , and a message asserting the conditional ‘if  $B$  then  $C$ ’. Then

$$B_k I_s^* B \wedge B_k I_s^* (B \rightarrow C) \tag{3}$$

Since the belief modality is normal it follows that

$$B_k (I_s^* B \wedge I_s^* (B \rightarrow C)) \tag{4}$$

Since the  $I_s^*$  modality is also normal, we also have

$$\vdash (I_s^* B \wedge I_s^* (B \rightarrow C)) \rightarrow I_s^* C \tag{5}$$

Since the belief modality, as a normal modality, is closed under logical consequence, it now follows from (4) and (5) that

$$B_k I_s^* C \tag{6}$$

Suppose now that, prior to receiving the two assertions,  $k$  already had the belief that  $C$  is false, i.e.,  $B_k \neg C$ . Since the D. schema holds for the belief modality, it now follows that  $\neg B_k C$ , from which it follows by the normality of the belief modality that

$$\neg (B_k B \wedge B_k (B \rightarrow C)) \tag{7}$$

From this it now follows that  $k$  cannot trust both of the messages he has received, so long as he retains his belief (which he might, of course, choose to revise) that  $C$  is false. This is a rather simple example, but it nevertheless serves to exhibit how the combination of the logics of the belief and ideality operators may be used to represent aspects of the recipient  $k$ 's reasoning, as he tries to work out which messages he can trust.

The speech-act theory literature has not paid a great deal of attention to the *reliability* of communicative acts; by contrast, however, issues pertaining to the *sincerity* of the communicator have figured prominently. There is, of course, an associated notion of trust here too. In terms of the general form of communicative conventions, as expressed by (sc-assert), we may say that an agent  $k$  trusts the *sincerity* of agent  $j$ 's communicative act if, having witnessed  $j$ 's performance of the act  $E_j A$ , and having then formed a belief of type (1), he then goes on to form the belief expressed by:

$$B_k B_j B \tag{2}$$

Cases in which an agent trusts both the reliability and the sincerity of a communicative act are commonplace. Similarly, lack of trust in reliability often goes hand-in-hand with lack of trust in sincerity. Furthermore, we may also have cases in which an agent trusts the sincerity of a communicator, but not his reliability—the communicator  $j$  is deemed to be describing the situation to the best of his knowledge and belief, but the source of  $j$ 's information is deemed to be unreliable. Finally, and less frequently, there may be cases in which communicator  $j$  is deemed by the audience to be insincere, and yet the audience trusts that the message he sends is reliable: the audience rightly thinks that the

communicator is trying to deceive them, but—unbeknown to the communicator—the information he is transmitting is known by the audience to be from a reliable source.

The distinction between trust in reliability and trust in sincerity also affords a means of explaining Moore's problem about 'saying and disbelieving' [13, p. 125]. In essence the problem is that while the sentence

$$A \text{ but I do not believe that } A \quad (8)$$

is not a logical contradiction, there is nevertheless something logically odd about it, and it is the nature of this logical oddity that needs to be explained.

Consider an act of uttering sentence (8) from the point of view of an audience  $k$ , and suppose that the utterer is (correctly) believed by  $k$  to be  $j$ . Suppose further that  $k$  trusts the reliability of the utterance. Then

$$B_k(A \wedge \neg B_j A) \quad (9)$$

from which it follows by elementary properties of the logic of belief, as a normal modality, that

$$B_k \neg B_j A \quad (10)$$

And now let us also assume that  $k$  trusts the sincerity of  $j$ 's utterance. Then

$$B_k B_j (A \wedge \neg B_j A) \quad (11)$$

from which it follows, again by elementary properties of the logic of belief, that

$$B_k B_j A \quad (12)$$

Note that (9) itself is logically consistent. And, given that we do not adopt the positive introspection axiom for the logic of belief, so is (11). However, (10) and (12) together represent incompatible beliefs of the agent  $k$ —logically incompatible beliefs if, as is usual, the D. schema holds for the logic of belief.

Thus, in relation to the previous discussion of the two types of audience trust, the diagnosis of the Moore problem is this: in contrast with most ordinary utterances, the two types of trusting attitude are here incompatible. The audience can trust the reliability of an utterance of (8) if and only if he does not trust its sincerity. (Equivalently, trust in sincerity is possible if and only if the utterance is deemed unreliable.)

We see two advantages of this explanation of the puzzle, as compared to that offered by Hintikka ([14, §§4.5–4.7]). First, our diagnosis explains the defective nature of the communication of (8), rather than considering (8) merely from the point of view of what the agent referred to by 'I' could himself believe about (8). Secondly, Hintikka's explanation of the logical oddity, unlike ours, turns essentially on his acceptance of the positive introspection axiom (the schema 4 in the Chellas classification) for the logic of belief. In our opinion, the informally stated explanation of the puzzle offered by Searle [7, p. 65, Footnote 1] is essentially the same as Hintikka's. Its focus is on the communicator's sincerity, and does not make explicit the tension between sincerity and reliability which we see as the key to Moore's puzzle.

## 5. Business communication and the Trusted Third Party

In [5] and [15] we develop a synthesis of Jones's convention-based analysis of communicative acts and Kimbrough's FLBC (Formal Language for Business Communication, see [16–22]), together with a detailed look at how the resulting combined formal models might be applied to the description of a trading scenario, involving, essentially, a buyer, a seller and a TTP (Trusted Third Party). Both [5] and [23] also discuss design of a Prolog implementation of the combined model. This combined model affords the prospect of deep and, we believe, plausibly complete formal integration of the theory described in this paper with the mundane, but complex, requirements of modern transaction processing. Moreover, we believe that the combined model will facilitate, in an entirely practicable and deployable manner, automated reasoning about communicated messages. These claims are under development and investigation. We content ourselves here with a brief indication of how the notions of conventional signalling systems, discussed in this paper, may be extended to support reasoning with additional sources of information.

Consider, then, a scenario in which a seller of goods and a prospective buyer communicate with each other not directly, but via a TTP. The seller,  $v$ , and the buyer,  $b$ , send via TTP messages of various kinds, which will typically

include (among others) messages that serve to state facts about available goods and their mode of delivery, to request information, and—if a deal is initiated—to create obligations. As the recipient of these messages, the TTP (agent  $t$ ) forms a set of beliefs of the type exhibited by

$$B_t I_s^* B \quad (\text{cf. (1), above}) \quad (13)$$

where, as we have earlier emphasised, the scope formula to the right of the  $I_s^*$  operator may take a number of different forms, depending on the nature of the communicative act performed.

In terms of our formal theory, the role of the Third Party, qua *Trusted* Third Party, is easily articulated. The key task for which TTP is responsible is to determine which inferences may be accepted from schemas of type (13) to schemas of type

$$B_t B \quad (\text{cf. (2), above}) \quad (14)$$

and then to communicate to buyer and seller the result of his deliberations. Since  $t$  is assumed by  $v$  and  $b$  to be trusted, they will accept what he says as true (they may even be obligated to do so by the contractual agreements they made in order to participate in the system). In other words, for  $v$  and  $b$  the task of making inferences from schemas of type (1) to schemas of type (2) has been delegated to  $t$ : they trust him to do that job for them.

Read schemas of the form  $Says_t B/A$  as ‘ $t$  says that  $B$  by seeing to it that  $A$ ’, where it is understood that ‘says’ is a generic term, referring to any type of communicative act. We define  $Says_t B/A$  as follows

$$(\text{Df. says}) \quad Says_t B/A \stackrel{\text{def}}{=} (E_t A \wedge (E_t A \Rightarrow_s I_s^* B))$$

where, as before, it is understood (i) that  $s$  is the conventional signalling (or communication) system that the agents  $t$ ,  $v$  and  $b$  have adopted (for the purposes of their trade communication) and (ii) that the scope formula  $B$  may exhibit a range of different forms, depending on which type of communicative act  $E_t A$  is.

Then we may represent the trusting beliefs that  $v$  and  $b$  have, vis-à-vis  $t$ , in the following way

$$B_v (Says_t B/A \rightarrow B) \quad (15)$$

$$B_b (Says_t B/A \rightarrow B) \quad (16)$$

And we may also wish to add that  $v$  and  $b$  and  $t$  are mutually aware that  $v$  and  $b$  have these trusting beliefs.<sup>13</sup>

We might say that schemas (15) and (16) articulate the *reliability-policy* adopted by  $v$  and  $b$ . That is to say, this is the policy they implement in order to solve the *reliability problem*: the problem of deciding when to make inferences from schemas of type (1) to schemas of type (2). In the absence of reliability-policies, it is clear that key aspects of the day-to-day operations of organisations—and indeed of interpersonal interaction in less formal settings—would break down, for the obvious reason that the agents concerned would not know which messages they should trust.

Using the services of a TTP is just one way of dealing with the reliability problem, just one type of basis for a reliability-policy, and it is presumably particularly useful in contexts in which communicating agents have no experience of each other’s behaviour, such as in a ‘first-trade’ scenario. Moreover, even in the presence of a TTP it is easily the case that not all aspects of what is communicated will be warranted by the TTP. For example, TTP may warrant a claim by the seller that the goods shipped have been insured, but not warrant any claim by the seller as to the contents of the materials shipped. The duties of a TTP are normally circumscribed. In these and other contexts it may be more sensible to adopt quite a different type of policy, for instance a policy that pertains to what is known about the reputation of the agents involved, their previous history with respect to reliability. Alternatively, in the absence of a TTP and in the absence of reputation indicators, the policy might refer to control and/or insurance mechanisms. Then an agent would accept the truth of a message on the grounds that he believes that the expected consequences for the communicator of unreliability would effectively deter him from transmitting a falsehood. Or the agent would trust a message on the grounds that—even if it did turn out to be false—he would run no real risk since he would be adequately protected by insurance.

At the extreme, perhaps, an agent might put into place reliability-policies effecting inference from schemas of type (1) to schemas of type (2) *even when the agent has good reason to believe that  $B$  is false*. For example, let  $B$  be the

<sup>13</sup> For an outline account of a logic of mutual belief, see [4].

seller's assertion that a given order has been fulfilled. Suppose that the order was for 2000 widgets exactly, but the number that actually were delivered was a little different, either higher or lower. The buyer may well want to have a policy to accept the order as fulfilled, given the seller's record (and the cost of making the necessary adjustments). The buyer might then declare the order fulfilled for the sake of accounting and payments purposes, while at the same time record accurately for inventory purposes the actual number. Such are everyday, unavoidable occurrences in commerce (although recording accurately for inventory purposes may be the exception!). Details, of course, need to be articulated, investigated and explored, but the potential for the present formal framework to accommodate these kinds of situations is indeed a significant virtue.<sup>14</sup>

For the sake of seeing the generalisation, it may be helpful to articulate the present example formally. More carefully, then, assume that the buyer,  $b$ , has ordered exactly 2000 widgets from the vendor,  $v$ . The vendor responds by shipping a number of widgets modestly different from 2000. (Whether it is higher or lower is immaterial for this example. In practice it may be either.) The vendor also sends a message (e.g., in the form of a packing slip) asserting that the package it shipped contains exactly 2000 widgets. The buyer,  $b$ , or an agent (e.g., employee) for the buyer,  $b_a$ , is at the loading dock when the package arrives. The agent's job is to inspect the package and to decide whether or not to accept it. For the sake of the example, let us assume that the agent correctly and accurately inspects the package, discerning that 2000 widgets were ordered, that it is claimed that 2000 widgets are delivered, and that in fact a number other than 2000 widgets were delivered. Let us also assume, for the sake of the example, that the variance of  $-2$  (under-count from 2000) is, according to the buyer's policies, within the margin of tolerance for paying the vendor but not within the margin of tolerance for recording actual inventory on hand. Our agent now has something of a dilemma. If the vendor is to be paid, the number of widgets must be declared to be 2000. If the inventory system is to be kept accurately, the number of widgets must be declared accurately, 1998.

A straightforward solution is for the agent to make two declarations. First, let  $C(2000)$  stand for the content of  $v$ 's utterance, viz., that "there are exactly 2000 widgets delivered." Let  $A1$  stand for this utterance (roughly " $v$  transmits the sentence 'there are exactly 2000 widgets delivered'"). Now, by (sc-assert) we have:

$$E_v A1 \Rightarrow_s I_s^* C(2000) \tag{17}$$

The buyer's agent,  $b_a$ , sees things differently and will *declare* to the inventory system that 1998 widgets have been delivered. Using  $A2$  for " $b_a$  transmits the sentence '1998 widgets have been delivered'" we have:

$$E_{b_a} A2 \Rightarrow_s I_s^* E_{b_a} C(1998) \tag{18}$$

The agent does this by, say, recording in the inventory system records the addition of 1998 widgets. What is recorded, typically, is that 1998 widgets are added to the inventory of widgets *and* that agent  $b_a$  did the adding (saw to it that the widget records were added). Subsequently, a user of the inventory system,  $x$ , may (typically would) believe that 1998 widgets were in fact delivered (and added to inventory)—

$$B_x C(1998) \tag{19}$$

—because  $x$  would trust that the ideality conditions were satisfied. The agent, on this analysis, would also record a second declaration. Letting  $AC(2000)$  stand for "there are acceptably close to 2000 widgets delivered," and using  $A3$  for " $b_a$  transmits the sentence 'acceptably close to 2000 widgets are delivered'" the agent would declare this to be the case, viz.

$$E_{b_a} A3 \Rightarrow_s I_s^* E_{b_a} AC(2000) \tag{20}$$

Points arising:

1. In each case, (18) and (20), the antecedents are modifications of a record keeping system. Typically, the agent creates or modifies computerised records of accounts. The agent is (ideally) empowered to make these changes (that's a major part of the agent's job).

<sup>14</sup> And of course it is well known that not all reliability-policies are sensible, wise or rational. For instance, recent history is littered with examples of politicians who have been trusted because their media image conveyed an impression of sincerity: the oft-trodden path of gullible's travels.

2. The analysis here correctly (we submit) identifies what, ideally, is the case in consequence of these actions by the agent. Ideally the agent has seen to it that 1998 real widgets have been added to the warehouse inventory. Further, ideally the agent has been diligent and honest, and in fact any variance from the contracted amount is negligible according to company policy.
3. The two declarations can go wrong in different ways. Regarding (18), the agent might mistakenly count the number of widgets, or lose them between the loading dock and the warehouse, or record the amount negligently or dishonestly. Regarding (20), the agent might make the declaration mistakenly, or without attending to its accuracy, or falsely by being bribed, and so on.
4. It is not the business of logic to prevent abuses of conventions or shortfalls from ideality. Logic can, however, contribute by affording rigorous representation and accompanying clarity. Future research will be required to articulate fully the conditions that suffice to establish or defeat ideality and the plausibility of ideality. Establishing that is an important task has been one of the goals of this paper.

Finally, with respect to the formal framework we have outlined in this paper, the key point to note is that it is not in any way tied to any special assumptions about the form or content of what we are here calling ‘reliability-policies’. Given our earlier critical observations about such Grice-inspired approaches as that taken by FIPA, it is particularly important that, on our account, reliability-policies need not presuppose that, to be trusted, communicating agents must have certain types of mental states or intentions.

## Acknowledgements

Much of the research reported here has been carried out within the EU project ALFEBIITE (IST-1999-10298), the EU Working Group iTRUST (IST-2001-34910) and the EU 6th Framework Integrated Project TrustCoM (<http://www.eu-trust-com.com>). The financial support of the EU is gratefully acknowledged, as is one of the JAL reviewers of this paper, who provided a number of very useful comments. This paper is a revised and expanded version of [25]. Material contained in [25] is here reproduced with kind permission of Springer Science and Business Media.

## References

- [1] M. Verdicchio, M. Colombetti, A logical model of social commitment for agent communication, in: F. Dignum (Ed.), *Advances in Agent Communication*, in: *Lecture Notes in Computer Science*, vol. 2922, Springer, Berlin, Heidelberg, New York, 2004, pp. 128–145.
- [2] M.P. Singh, Agent communication languages: Rethinking the principles, *IEEE Computer* 31 (12) (1998) 40–47.
- [3] A.J. Jones, X. Parent, Conventional signalling acts and conversation, in: F. Dignum (Ed.), *Advances in Agent Communication*, in: *Lecture Notes in Computer Science*, vol. 2922, Springer, Berlin, Heidelberg, New York, 2004, pp. 1–17.
- [4] A.J. Jones, X. Parent, A convention-based approach to agent communication languages, *Group Decision and Negotiation* 16 (2007) 101–141. Available online at <http://www.springerlink.com/content/wn13u56428405345/>.
- [5] A.J. Jones, S.O. Kimbrough, A note on modelling speech acts as signalling conventions, in: S.O. Kimbrough, D.J. Wu (Eds.), *Formal Modelling in Electronic Commerce*, in: *International Handbooks on Information Systems*, Springer, Berlin, 2005, pp. 325–342.
- [6] E. Stenius, Mood and language game, *Synthese* 17 (1967) 254–274.
- [7] J.R. Searle, *Speech Acts*, Cambridge University Press, Cambridge, 1969.
- [8] A.J. Jones, *Communication and Meaning—An Essay in Applied Modal Logic*, Synthese Library, vol. 168, D. Reidel, Dordrecht, 1983.
- [9] A.J. Jones, M.J. Sergot, A formal characterisation of institutionalised power, *Journal of the Interest Group in Pure and Applied Logic (IGPL)* 4 (3) (1996) 427–443. Reprinted in [24, pp. 349–367].
- [10] A.J. Jones, On normative-informational positions, in: A. Lomuscio, D. Nute (Eds.), *Deontic Logic in Computer Science*, Proceedings of the 7th Int Workshop on Deontic Logic in Computer Science, DEON 2004, Springer, Berlin, 2004, pp. 182–190.
- [11] J.R. Searle, D. Vanderveken, *Foundations of Illocutionary Logic*, Cambridge University Press, Cambridge, 1985.
- [12] J.L. Austin, *How to Do Things with Words*, Oxford at the Clarendon Press, Oxford, England, 1962.
- [13] G.E. Moore, *Ethics*, Home University Library, London, UK, 1912.
- [14] J. Hintikka, *Knowledge and Belief—An Introduction to the Logic of the Two Notions*, Cornell University Press, Ithaca, NY, 1962.
- [15] A.J. Jones, S.O. Kimbrough, A convention-based approach to a formal language for business communication, Draft manuscript, University of Pennsylvania, Philadelphia, PA, 2006.
- [16] S.O. Kimbrough, EDI, XML, and the transparency problem in electronic commerce, in: S.O. Kimbrough, D.J. Wu (Eds.), *Formal Modelling in Electronic Commerce*, in: *International Handbooks on Information Systems*, Springer, Berlin, 2005, pp. 201–227.
- [17] S.O. Kimbrough, Y.H. Tan, On lean messaging with unfolding and unwrapping for electronic commerce, *International Journal of Electronic Commerce* 5 (1) (2000) 83–108.

- [18] S.O. Kimbrough, S.A. Moore, On automated message processing in electronic commerce and work support systems: Speech act theory and expressive felicity, *ACM Transactions on Information Systems* 15 (4) (October 1997) 321–367.
- [19] S.O. Kimbrough, Reasoning about the objects of attitudes and operators: Towards a disquotation theory for representation of propositional content, in: *Proceedings of ICAIL '01, International Conference on Artificial Intelligence and Law*, 2001.
- [20] S.O. Kimbrough, Y. Yang, On representing special languages with FLBC: Message markers and reference fixing in SeaSpeak, in: S.O. Kimbrough, D.J. Wu (Eds.), *Formal Modelling in Electronic Commerce*, in: *International Handbooks on Information Systems*, ISBN 3-540-21431-3, Springer, Berlin, 2005, pp. 297–324.
- [21] S.O. Kimbrough, A note on interpretations for federated languages and the use of disquotation, in: A. Gardner (Ed.), *Proceedings of the Tenth International Conference on Artificial Intelligence and Law (ICAIL-2005)*, Bologna, Italy, 2005, pp. 10–19; In cooperation with ACM SIGART and The American Association for Artificial Intelligence.
- [22] S.O. Kimbrough, A note on the Good Samaritan paradox and the disquotation theory of propositional content, in: J. Horty, A.J. Jones (Eds.), *Proceedings of ΔEON'02, Sixth International Workshop on Deontic Logic in Computer Science*, May 2002, pp. 139–148.
- [23] A.S. Abrahams, J.M. Bacon, D.M. Evers, A.J. Jones, S.O. Kimbrough, Introducing the fair and logical trade project, in: *Workshop on Contract Architectures and Languages (CoALa2005)*, 2005.
- [24] E.G. Valdés, et al. (Eds.), *Normative Systems in Legal and Moral Theory—Festschrift for Carlos E. Alchourrón and Eugenio Bulygin*, Duncker & Humblot, Berlin, 1997.
- [25] A.J. Jones, S.O. Kimbrough, On the normative aspect of signalling conventions, in: L. Goble, J.-J.C. Meyer (Eds.), *Deontic Logic and Artificial Normative Systems*, *Proceedings of the 8th International Workshop on Deontic Logic in Computer Science, DEON 2006*, in: *LNAI*, vol. 4048, Springer, Berlin, 2006, pp. 149–160.