Meaning Negotiation and Communicative Rationality

R.A.Young
Philosophy Department
University of Dundee
r.a.young@dundee.ac.uk

Abstract

Settling disputes, even disputes about meaning, by third party arbitration is different from settling them by negotiation. Moreover, disputes can be settled by a procedure in which the disputing parties follow a protocol that prohibits them from simply following their original aims. They can collectively constitute an arbitrator. This provides an alternative model for resolving differences in meaning from simple negotiation. In his theory of communicative rationality, Habermas has a model in which human languages, in effect, invariably provide a system of collective arbitration. Indeed human agents may be best understood by treating their *aims* as derivative and their *needs* as fundamental. Arbitration can operate by respecting *needs* but not necessarily *aims*. In communicative rationality, needs may come to be recognised which an agent did not originally understand itself to have, so the agent's aims may change. Thus it is distinct from strategic (game-theoretical) or instrumental (means-end) rationality, because these presuppose fixed aims for each agent. Artificial agents may be constructible on this model, but it is suggested that they would be complex, and no blueprint is proposed. However, a formal account of the model is sketched.

Introduction

This paper considers a specific procedure for interactively reaching agreement on meaning. However, where there is a need to reach such agreement, it is typical that there are other needs to be fulfilled. Therefore the paper, despite its focus on meaning, will consider it in a broader context.

Consider an agent Ego that enters into a binding arbitration agreement with agent Alter to resolve a dispute between the pair of them. The dispute may be about the meaning of certain terms that are used to express rules which confer rights on the disputing parties. If the arbitrator interprets these terms in one way, then agent Ego's claims to rights are vindicated, and agent Ego can benefit in pursuing aims. If the decision goes the other way, then agent Alter will benefit. Also, the arbitrator might make a decision that benefits neither party. If the agents were human then this might be some set of lawyers, for example.

Once the arbitrator makes a decision then at least one of the two parties has some constraint placed on its action. Suppose Ego is constrained. Ego is bound by the decision and must respect it. Unless the decision specifically requires Ego to give up some aim, or to weight aims differently, the constraint will restrict the means that Ego may use to pursue aims. It will be what has been called a side-constraint (Nozick 1974, p.29–35). Suppose first of all that, before the arbitration, Ego had no such side-constraints[1] but was completely free to pursue its aims. Now, after the arbitration, Ego enters into a new form of decision-making about action. Ego no longer simply identifies the best means for attaining ends, because Ego is bound to reject any means that have been ruled out by the arbitrator's decision.

It is usual to distinguish negotiation, in which the negotiating parties strive to reach agreement without the intervention of an arbitrator, from third-party arbitration. In direct negotiation, the parties preserve the unfettered power to decide on their course of action for themselves. Even so, in negotiation, some side-constraints on action may be recognised. Negotiation proceeds towards a contract. If a contract has been accepted only because of the deceit of one of the negotiators, then it may be deemed null and void. Thus a protocol for negotiation may rule out deceit. Nevertheless, negotiators need not provide for the aims of the other parties to the negotiation, unless it is in their interest to do so. Negotiators may be required to satisfy themselves that the negotiation sufficiently fulfils their aims; they have the responsibility and power to look after their own aims. Under arbitration, some of this responsibility and power has been surrendered to the arbitrator[2].

Why do I consider arbitration, when the topic of the workshop is meaning negotiation? It provides a route into making a distinction that is often left obscure. Consider what happens, if Ego and Alter do not appoint a third-party arbitrator, but instead enter into an interaction between the two

[1] In our original example, Ego's dispute with Alter was about rights, and it is arguable that, if one recognises any rights, then one recognises side-constraints on action. In any case in this paper, we will consider agents for which a system of rights (expressed in side-constraints) is presupposed from the beginning

[2] A complication is that there can be cases of indirect negotiation where one or more parties appoints an agent (for example a lawyer) to negotiate on its behalf. Nonetheless, this is different from arbitration, because the new negotiating agent applies the power of one of the parties on its behalf, whereas the arbitrator does not necessarily act on behalf of any one party.

of them according to a special protocol which is designed to achieve a final binding decision on their dispute, Suppose the special protocol invokes side-constraints that prohibit each agent from simply following its own original aims. In that case, when the two agents follow the protocol, surely they constitute an arbitrator. Therefore they are not engaged in negotiation, but in a form of collective deliberation that is tantamount to arbitration. It is often obscure in discussions employing the term 'meaning negotiation' whether or not the envisaged interaction excludes collective arbitration. It is important to clarify this, because full-blooded negotiation, in which each agent unscrupulously follows its own aims, is a very different model for deliberating about meaning from interaction according to a protocol that constitutes collective arbitration.

What I shall explore in this paper is Habermas' conception of communicative rationality (Habermas 1998). We may use collective arbitration as an example of it. Following Austin's (Austin 1975) account of illocutionary acts, Habermas argues that in human linguistic practices we invariable engage in linguistic acts that imply side-constraints on action. Thus our linguistic activity is, according to him, not simply intelligible as action in which each agent simply pursues its own aims. Also, according to Habermas, in human linguistic practice we invariably have the possibility of engaging in action that is, at least when there is a dispute, tantamount to collective arbitration. This is how we resolve disputes about meaning, if we act in accordance with communicative rationality. There are also cases when there is no dispute, but nevertheless meaning is unclear. In that case too, we can employ communicative rationality.

Aims and Needs

In some cases, agents may be fortunate enough to begin with aims that are not incompatible. In that case, one might think, they may have problems coordinating, but do not have conflict. However a problem arises if each agent understands its aims through ambiguous language. There may be potential conflict, if the ambiguity of the various descriptions of aims is resolved antagonistically, but it may be possible to resolve ambiguities in a way that renders them compatible. The worst problem arises if aims are straightforwardly incompatible.

In the case where one set of disambiguations can avoid conflict, whereas another can create it, one can see that there might be a role for arbitration. If, in the end, arbitration produces a solution that enables each agent to implement the arbitration in a way that is not in conflict with its disambiguated aims, then the arbitrator can provide an unproblematic solution. Indeed, in this case, an arbitrator may not be necessary. Instead mediation (the mediator provides disambiguations which the parties, left to themselves, would not have identified) may be all that is necessary. If agents are able to follow a special protocol, then it may even be possible for there to be a collective mediator capable of discovering disambiguations that agents directly attempting to pursue their aims would not discover.

However, what is the role of arbitration, if the aims of the agents are straightforwardly incompatible? Suppose agents have the ability bindingly to commit themselves to apply side-constraints if the arbitrator requires it. In that case, agents will make commitments to arbitration in ignorance of its result, each hoping for favourable arbitration. In the event that arbitration goes against an agent, it is left with a binding commitment that conflicts with its original aims. There can be an external mechanism of binding commitment, in which agents agree to external sanctions being applied to them if they fail to act on their commitment. In that event, in a sense, the aims of the agents cease to be in conflict with the commitment, because of sanctions. As well as external mechanisms, one can envisage internal mechanisms that enable binding commitment. One mechanism is to have internal sanctions, in humans, perhaps the threat of guilt. However, one can envisage less complicated mechanisms, for an agent may simply have the ability to apply rules that the arbitrator ordains. For example, perhaps humans can act on duty, overriding their own original aims.

If original aims are overridden, does this mean that there is a change in aims? If aims are ephemeral in this way, then what is the fundamental basis for justifying action? Suppose, in the human case, one thought that the fundamental mechanism was that of binding commitments that could not be retracted. In that case, a human agent could undertake a binding commitment to some course of action, and there would be no rational route back from it. In human affairs, this way of thinking can have tragic consequences. People can commit atrocities out of what they take to be binding duty.

A quite different way of thinking is that it is not aims that are fundamental, but needs. Aims are practical premises for reasoning that we recognise and implement in plans. We may in some cases fail to recognise our aims explicitly, but in those cases they are implicit in our actions. They are implied premises, and our actions implement them. In contrast, we may entirely fail to recognise our own needs. To the extent that we are rational, we may have the capacity to recognise needs, but it does not follow that we actually recognise them.

If we look beyond the purely linguistic component of Habermas' theorising, then we have an account of needs (Habermas 1979, p78, pp.90–91) that may eventually be articulated and recognised in language. Thus, if communicative rationality provides a procedure tantamount to collective arbitration, that arbitration may take effect, not through the application of sanctions or of mere binding commitment, but through the recognition of the needs of each person. Communicative rationality operates by people coming to assent to an account of their needs that enables them to recognise commitments and to act on them. However, commitments are not irrevocable, because arguments may be presented to show that needs override them.

How are needs to be identified? Sometimes we speak of needs relative to aims, "If you aim to go to the airport, then what you need to do is ...". However, if we are treating needs as fundamental, not aims, then this cannot be our account of fundamental needs. Instead, as beings engaged in communication, we may identify needs as those human requirements which we can agree, in rational communication,

to recognise.

Agents

What kind of agent can engage in communicative rationality? Consider two factors that have been introduced in thinking about communicative it:

- side-constraints
- needs which an agent has, but which it may not recognise itself to have

In this section, we shall consider how these factors might be introduced into a belief-desire-intention model (Bratman 1987). As well as providing for these factors, we obviously need multiple agents with a capacity to communicate with each other. We also need to provide for change in the meaning of the language in which communication takes place, since this is an important aspect of communicative rationality, and in any case it is the topic of our workshop. However, let us first consider side-constraints and needs.

If we consider a set of aims or desires, and these have a priority ordering, then we can understand how side-constraints might apply to actions on low-priority aims. The side-constraints might be necessary in order to secure the fulfilment of higher-priority aims. However, when I considered arbitration, I did not suggest that the commitment to the side-constraint proposed by the arbitrator was to be contingent on the commitment meeting some high-priority aim. Instead the side-constraint was to apply to action on all aims. Indeed it was suggested that certain side-constraints might be implied in language itself. Later, it was argued that side-constraints might be revised in recognition of needs.

In a belief-desire-intention model side-constraints might be introduced as contents of intentions, but if they are to be capable of constraining action on *all* aims, then the status of no aim can be so important that all intentions are conditional on its fulfilment. Thus the status of aims is revisable. At least to the extent that they can be given a lower priority by the formation of intentions to fulfil side-constraints. Also, it would seem that, if an agent comes to recognise a need that has hitherto been unrecognised and not acted on, then the agent can come to have new aims. We may also consider that rational aims or desires will be abandoned if it is recognised that their fulfilment is impossible. Thus we might think of aims as highly revisable, as items which can come and go much like intentions.

If aims are to come and go, then how is this to be possible? One way in which an agent might come to have a new aim is that it might be proposed by another agent. Yet, if the recipient of the proposal is to adopt such an aim rationally, then there must be some constraint on the aims that it can adopt. The side-constraints that it already intends to respect constitute a constraint, but it has been argued that, in humans at any rate, such side-constraints can be revised in the light of needs. Are there other constraints? A constraint that we have just recognised is that an aim may be rejected if its fulfilment is impossible for the agent to carry it out. In humans this may come about because the agent is incapacitated by some emotional process. Thus a person might plan to learn to rock-climb and yet, when the lesson begins, be incapacitated by fear. At least for a time, the person may need to recognise their own incapacity. Thus we can think that there are semi-autonomous processes within a person that can incapacitate action or alternatively facilitate, as fear may facilitate flight.

In an artificial agent, such processes might reduce the resources available for carrying out a plan (as fear may incapacitate, or at least render difficult, movement into danger) or enhance them (as fear may facilitate flight). If an artificial agent has such semi-autonomous processes, and a theory of how they operate, then its reasoning about a proposed aim can in part consist in assessing whether it is feasible given the constraints provided by these semi-autonomous processes. Thus we might have agents that are capable of assenting to plans that are proposed to them, even though these plans conflict with established ones, indeed with premises that they have hitherto recognised in practical reasoning.

Thus a person might plan to avoid rock climbing, because it is felt to be too dangerous. Another might plan to avoid flying in aeroplanes, for the same reason. In these cases, our fear, our sense of danger, indeed our understanding of the term 'danger' is malleable. With sufficient training, we may come to be able to perform, indeed we may even enjoy, actions, for which we previously felt incapacitating fear. Yet there are arguments about whether it is always rational to allow fear to be desensitised. Case by case, we can argue with each other, about what we need by way of a sense of fear. In the human case, we do not just have fear, but anger, pride and much else. In each case, these are malleable, and we can argue about their proper role. If we are to think of artificial agents as having semi-autonomous sub-processes akin to fear, then we need to think of those processes as changeable by their own learning algorithms, perhaps connectionist ones. In a human, fear does not simply disappear through reasoning, but through training.

In a design for an artificial agent capable of deliberating about needs, as opposed to aims, through communicative rationality, it might be appropriate to specify its program in a non-deterministic way. It would have a set of semi-autonomous sub-processes operating in parallel. Its reasoning processes might also operate in parallel. It is possible to specify parallel programs permissively, recognising that their outcome may be dependent upon the exact time sequence in which processes interact and demand resources. Thus we need not think of am agent's preferences as being completely determined by the specification of its system. The system might be deterministic at a physical level, but it would not follow from this that a preference ordering was determined, because the outcome of its decision making might be dependent on fine details of its interaction with the external world, including other agents, and the details of ensuing calls on resources. It is not that the agent would have preferences about these fine details, but just that the fine details would vary its decision making. Perhaps this consideration complicates the design of agents, but perhaps artificial agents designed in this way would be human-like. There would be no simple definition of utility for agents of this kind. If agents are flexible with respect to preference orderings, and therefore utility, but capable of committing

to intentions and therefore contracts, then they may be better able to achieve agreement than agents with determinate (and internally rationally consistent) preference orderings, but whose preference orderings are incompatible with those of other agents.

What would constitute a *need* for an agent as malleable, as capable of learning, as a human? The Habermas' view (Habermas 1979, p.78,pp.90–91) is that a rational agent needs something if and only if (1) that thing is required to fulfil an aim that the agent would pursue after sufficient rational communication and (2) the agent would not cease to pursue that aim even if rational communication continued indefinitely under conditions that facilitated learning. This is given the background of rational constraints on the individuals involved, including rational constraints of communication. Amongst the needs that are to be defined in this way are needs to use language for reference, for description, for contract, and so on. Plainly an important question about his position is whether our problem-solving skills, even when they are combined in rational discussion among agents, are adequate to the task. To be sure, at least in the way that I have defined it, a need might never be explicitly recognised by an agent, even under the best practicable conditions of communication and learning. It might only be implied by an aim that the agent would recognise under these conditions. Thus the definition allows that there might solutions, and therefore needs, that are beyond the agents themselves. Nevertheless these needs would themselves be relative to aims that the agents are capable of recognising.

To be comparable with a human, an artificial agent capable of engaging in communicative rationality would need to be sufficiently complex to combine all the relevant capacities for communication and planning with motivational subagents capable of learning. Whilst I think that such agents may be feasible, I do not purport, in this paper, to provide a blueprint them. However, in designing artificial agents, I do propose that we should recognise the full complexity of the human case, and then decide how we might abstract from it, or decide to design something different. In this paper, I discuss the human case, and then attempt to provide a formal model for its analysis.

The Workshop Example

In the announcement of this workshop, it is stated that:

> The idea of MN [meaning negotiation] is that any real-world approach to semantic interoperability between autonomous entities (namely entities that cannot assess semantic problems "by looking into each other's head", like humans or software agents) should involve (among other things) a social process of negotiating an agreement on the content (semantics) and the speaker's intention (pragmatics) of a communication.

This passage is very much open to dispute if it marks a contrast between an agent's understanding of its own meanings (in which it can tell what its meanings are by 'looking into its own head'), and of the meanings of others.

Take the example, from the announcement, of booking a holiday. Consider a Scot seeking a holiday on the 'green fringes' of the 'Mediterranean' countries. The Scot has browsed brochures on 'Green' Spain and 'Green' Italy. Does the Scot know exactly what he means by 'green' and 'fringe'? Asked what 'green' means he might reply that a 'green' place will have vegetation *sufficiently like* Scotland, and *some* refreshing rainfall and *enough* clouds to stop the sky being boringly blue, whilst still having some sunshine. This is not a clear account of the Scot's borderlines, and he may be less clear about them than the travel agent, who might already have had similar customers. What about the 'fringes' of the Mediterranean, does 'green' Portugal count — and what about Austria, if the Trentino counts as 'green' Italy, then why not count southern Austria? In practise, to settle the Scot's borderlines we need to understand his needs — does he need to have warmth without sunburn, or the Mediterranean diet and/or red wine (that rules out Austria) or mountains and lakes (that rules out Portugal)? Typically the travel agent will suggest holidays and the Scot will reject them or accept them provisionally whereupon price and availability will be checked, and very likely there will be more iterations.

In any exchange like this, the Scot reacts case-by-case and so in a sense 'knows' both 'what he means' and 'what he wants', but it is very misleading to think that the Scot has a systematic understanding of the intensions of the terms by which he categorises or of his preferences or aims or needs. He is like a naive user of a grammar, who distinguishes grammatical sentences case-by-case, but is unable to produce a useful theory of his grammatical rules. Note that a software agent may also be in this position, its program may enable it to use words to express categories or make choices without enabling it to know the rules that govern its use, or its ultimate needs.

Thus the Scot deliberates with the travel agent about what holiday to choose, what he might mean by 'green fringes' and what he wants in and from his holiday. He does this all at once, and a good travel agent, interested in keeping his business in the long term, facilitates this. It would be very misleading to model their discussion as a straightforward negotiation between agents each of whom understand his/her own needs and meanings thoroughly. Within this deliberation, there is a need to have a language that gives sufficient expression both to the Scot's needs and also the travel agent's business needs.

A further point is the stability of the Scot's evaluative and linguistic attitudes. Suppose the Scot is booking holidays on behalf of his family. Each of them might have a different understanding from the consensus developed between the Scot and the travel agent. Not only may the wife, children and widowed mother-in-law argue for different needs and meanings, they may even succeed in persuading the Scot to value and categorise differently. If the travel agent does indeed want a satisfied customer who is likely to return next year, then the travel agent needs to achieve a consensus on the meanings and needs of the customer that will be stable until the holiday is over, not just one that fits the agent's ephemeral attitudes months before the holiday takes place.

Habermas on Communicative Action

Habermas (1998) makes a distinction between kinds of rationality. He recognises what he calls (1) instrumental rationality, and also (2) strategic rationality, but he also insists that there is a distinct (3) communicative rationality. Communicative action is defined as action that is 'oriented toward understanding' and communicative rationality is the kind of rationality that pertains to it. Therefore communicative rationality is supposed to be pervasive in human linguistic practise, but it is especially manifest when there are disagreements about meaning and/or illocutionary force. (However, as we shall see, it may also have an important role in resolving other kinds of misunderstanding). Habermas' conception is that when we engage in dialogue beyond our normal linguistic practise we have to rely on general principles of communication.

The claim that there is a distinctive communicative rationality is pertinent to the present workshop. However, Habermas' work is far from formal. In the penultimate section of this paper, I sketch a formal approach. It is not claimed that this formal approach constitutes an exegesis of Habermas whose work is complicated and stimulating, but not formally precise.

Is Communicative Rationality distinct from Strategic Rationality?

In this line of thought, we need sharply to distinguish communicative action from strategic (game theoretic) action (Habermas 1998, p.118). If we list the three kinds of action, then we have:

Instrumental Action in which a single agent selects actions that serve some end(s), *this requires instrumental rationality, which identifies means to given ends*.

Strategic Action in which several agents need to solve problems of conflict and/or coordination in achieving their aims, *this requires strategic (game theoretic rationality) which utilises the capacity of each agent to reason out the strategies of the other agents*.

Communicative Action in which several agents engage in action that is 'oriented toward understanding' among the group, *this requires communicative rationality that operates within constraints that enable understanding and whose orientation is toward consensus on the discursive justifiability 'validity' of claims about linguistic acts, meanings and needs*

Note that, I have defined communicative action in terms of understanding among a group. Habermas himself thinks that, in communicative action, we aspire to action that is justifiable to any rational agent that interacts with the group under the constraints of communicative rationality. Thus, to be consistent with Habermas, my group would need to be a group open to any agent willing to accept the constraints.

A problem in distinguishing communicative from strategic rationality is that being 'oriented toward understanding' among a group might consist simply in having such understanding as one's dominant aim. In that case, the problem is a coordination problem among the group, or perhaps a combination of a coordination problem with a problem for instrumental reason if information needs to gained from the world. This point cannot simply be met by insisting that the constraints of communicative rationality are normative — they specify how we ought to act rather than how we invariably do act. Both instrumental and strategic rationality are also normative, and the question is whether the norms of communicative rationality can deduced from their norms in combination with the assumption that promoting understanding is the priority.

Nevertheless, communicative rationality can be distinguished from strategic rationality, if we take it that 'orientation toward understanding' is orientation toward understanding that includes self-understanding, because *amongst the things one does not (wholly) understand, but needs to understand, may be one's own needs themselves*. If one does not wholly understand one's needs, then one does not understand one's rational aims and preferences. Thus one has a problem that is not amenable to strategic reason, because strategic reason presupposes an understanding of one's rational aims. Thus, perhaps there is an opening for another kind of action, but there is a puzzle: *how is action intelligible at all, if one does not wholly understand one's needs and therefore one does not understand one's rational aims*?

To answer this question, we shall assume that any agent will have a basic understanding of the following kind: given a choice between two alternative actions under certain descriptions, and the assumption that fuller descriptions would prove neutral, an agent can understand what its current preference [3] is between these alternatives (even here there may be problems about its understanding all the implications of the two descriptions and also problems of the consistency of preferences). To be an agent it need not have a systematic understanding of its needs, rational preferences and aims, nor an understanding of how they will change over time as further information is gained.

Costs, benefits and power

In 'meaning negotiation' we are considering a change of meaning by some or all parties, which is brought about through their communication. In any such interaction each party needs to consider the costs and benefits of the change for it (including the cost of participating in the interaction that leads to change). One can analyse this from the point of view of strategic rationality if each parties' costs and benefits can be assessed in terms of its aims. However, within communicative rationality, costs and benefits are to be assessed relative to need. Questions of need are to be settled by discussion among the parties to the communication. In exploratory discussions, as we articulate the needs of the parties, the costs may prove unacceptable to one or more of the parties. It is unreasonable to expect any party to engage sincerely in communication if their participation is simply

[3] Earlier it was suggested that agents might be non-deterministic, and therefore not hadve determinate preference orderings. What is being considered here is an ephemeral preference on a particular occasion, not a knowledge of a fully determinate preference order.

to be exploited, at their cost, for the benefit of others. Thus if we are genuinely oriented towards communication we are constrained to ensure that no one will be exploited[4].

Of course, there are many interactions in which one or more parties do feel that it is within their power to exploit others. An employer may feel h/she has the power to exploit an employee, or a strong labour union a weak employer, or a monopolist a customer. In this case, the would-be exploiter cannot expect their intended exploitee to engage in sincere communication except through some threat of power or through deceit. The exploiter may purport to engage in communication, but it, together with lies or insincerities, is what Habermas (1998, p.93) would call manipulative or 'systematically distorted communication' (in this latter case the agent is involved in self-deception). In the case of manipulation, the exploiter itself should expect to lose the benefits of engaging in communicative rationality. It loses the benefit of recognition of its needs as opposed to its demands and it risks ultimate dissatisfaction in which it comes to understand that its demands were irrational because not related to its own needs. In the case of systematically distorted communication, the agent's interlocutors may hope that rational criticism will, in the end, overcome self-deception, but self-deception is complicated and some cases may require more than rational criticism. We will disregard it for present purposes.

Communicative Rationality

According to Habermas (1998, p.22), when I engage in communication I imply that my action meets certain constraints, indeed that I can vindicate it (justify it in argument) as meeting these constraints, always provided my interlocutors also operate within these constraints. I imply that my communicative action is:

1. intelligible (decipherable in some system of signs and having a meaning)
2. true (in the case of speech acts other than constatives, e.g. imperatives, there will still be implied presuppositions that need to be vindicated)
3. sincere (it expresses the attitudes someone is supposed to have when engaging in it)
4. right (justifiable as a contribution to communicative interaction and as appropriate to the needs of the parties.)

With respect to all these implied validity claims there are questions about what precisely their import is. In Habermas' view, claims about their exact import, much like the Scot's ways of discriminating holidays, stand or fall according to whether they can withstand cross questioning and argument. Thus what matters is whether they are sustainable in future action and communication. Thus the whole orientation, both of the theory of the pragmatics and of concrete cases, is toward understanding that is achieved in the future.

Yet it is part of the future-oriented process Habermas envisages that we do need to look backwards to past actions. A powerful reason for this is that communicative action will only achieve understanding if interlocutors approximate its constraints. Each interlocutor needs to ask whether each other interlocutor is respecting them. The Scot needs to ask whether the travel agent is trying to help him fulfil his needs, as opposed to (say) simply trying to sell him the most expensive holiday. The travel agent needs to ask whether the Scot is interested in a holiday or simply fantasising about holidays on a dreary winter evening.

Rational Reconstruction

When agents look backwards to their past actions there are invariably many interpretations or (more or less) rational reconstructions (Habermas 1998, p.29) of their behaviour especially if we allow for potential explanations in the sense of Nozick (1974, pp.7–8). Typically under some interpretations it will be sincere and benevolent, but under others it will be mendacious, dissembling or manipulative. In the form of communicative rationality that I shall propose each agent needs to be able to vindicate itself as an agent that has been attempting to fulfil the norms of communicative rationality, or at least as one that is now trying to fulfil them, and which has been learning, or is able to learn, to do so. A problem here is that there is potential for endless reinterpretation of what has happened and also of an agent's potential to learn. In practise, it is necessary to curtail this, otherwise no decision could ever be taken to deem a person to have flouted constraints so severely that they ought to be excluded from full communicative interaction. To remain within full communicative interaction, a concise explanation is needed (according to the group's currently preferred theory of agency) of how the agent does and will conform to the norms and of how this is consistent with its past behaviour.

Aims, Languages and Natures

To finish, I sketch a formal learning theory (Jain & al 1999; Kelly 1996; Young forthcoming) of communicative rationality. In formal learning theory, one proves, from a set of assumptions about a potentially infinite course of inquiry, that a particular method of pursuing the inquiry is reliable in the limit for meeting the requirements of the inquiry. Usually what is required is that an hypothesis be at least compatible with a data stream or perhaps predictive of it. In the case of communicative rationality, as well as requiring hypotheses that are predictive of human behaviour, we require that it produce expressions of norms and needs.

1. that are realistic in the sense of not requiring courses of action that are naturally impossible

[4] In the first section of this paper, it was stated that under protocols for negotiating a contract, agents might be required to tell the truth but not to ensure the fulfilment of the needs of other agents. Various protocols may be considered for contract. Thus, concealment of information about an item offered for purchase may by prohibited under a protocol for contract. Under communicative rationality, agents would be required to seek to fulfil the needs of others, and this, I think, goes beyond any normal conditions of contract. Therefore, it may be argued that Habermasian communicative rationality is over-idealised. We need to distinguish between contexts in which communicative rationality would be effective from ones in which it would not. In the penultimate section of this paper, an approach to this problem is sketched.

2. that vindicate the past activity of agents in a group of communicatively rational agents as justifiable or at least sufficiently excusable for them to continue as members of the group
3. that are acceptable in argument amongst the group
4. that constitute an expression of the norms of communicative rationality
5. that articulate what is needed in respect of compatibility of meaning and linguistic practises for communication to be feasible.

It is requirement 5 that pertains directly to meaning negotiation, but the main thrust of this paper is that we need to recognise that 'meaning negotiation' is embedded within the wider context of communicative rationality.

We shall say that sets of propositional attitudes combining hypotheses and expressions of norms and needs are vindicated against the data stream as meeting the requirements of communicative rationality if they meet the above requirements. In the case that they have been vindicated up to time t, we shall say that they are provisionally vindicated at time t. In the case that they are vindicated at t, and will continue to be vindicated if we do pursue inquiry indefinitely, we will say that they are vindicated against the actual data stream in the limit.

We can think of the natural world in which the agents are interacting as constituting one natural system. The natural system S consists of a tuple $\langle T, X, V_x, O, M\rangle$, where T is state transition relation between states, X is a set of state variables, V_x is a set of possible values for each $x \in X$, and O and M are subsets of X respectively representing the observable and the manipulable variables of the system. If s,r are states of the system, then $T(s,r)$ obtains if for each time step n, if the system is in s at n then it is possible[5] for the system to be in r at time step $n+1$. The sense of possibility is not epistemological but scientific (e.g. physical).

A normative requirement is placed upon this natural system. This is another relation N. If s is a state of the system and R is a set of states of the system, then $N(s,R)$ is satisfied if for each time step n, if the system is in s at n then the system is in some $r \in R$ at time step $n+1$. This normative requirement will, to the extent it is fulfilled, constrain agents to fulfil the requirements for communicative rationality. But this normative requirement is not conceived as simply a given. Instead it is a normative requirement that is to be compatible with the normative requirements that agents involved in communication would accept as justified by argument if their experience and interaction were continued indefinitely. The agents express their requirements linguistically, but N is the natural relation that must be instanced if their constraints are to be fulfilled.

[5]Thus, in this account a system is specified in a way that provides for non-determinism. The account is general enough to provide for determinism in which, at any given time, only one state transition would be possible. In the section on agents, earlier in this paper, non-determinism was considered as a possible feature of agents engaged in communicative interaction, and the formal analysis is designed to provide for it.

At any given time, an agent is in a context, where a context consists of the following tuple: $\langle K, i_0, i_t, S, N, \beta, H, L, U\rangle$, where K is a set of background assumptions, i_0 is the index (place and time) of commencement of the agent's activity, i_t is the index which the agent has reached, S is the system of the natural world from the index of commencement, N is the normative requirement,β is a method of inquiry, H is a preference ordering of sets of consistent propositional attitudes that the agent can take to the world (including hypotheses about the world, about agents and about languages, but also including normative attitudes), L is a pair of a language (states of S as its domain) and metalanguage, and U is a set of rules of inference, interpretation and argument. The assumptions K are expressible in L as axioms to which the rules U may be applied. Languages change as the agent recognises how exactly they need to change. For each $h \in H_g$ (where H_g is the ground set of the ordering H), there will be courses of inquiry $Inq(i_0, S, \beta, h)$ determined from the commencement of the context.

Each agent has a sequence of contexts through time. We can define relations of compatibility between agents and between agents and normative requirements. A formal learning theory within this framework would identify compatible sets of assumptions amongst a group of agents and a compatible normative requirement such that for each agent that fulfils the normative requirement, in its sequence of contexts, a set of propositional attitudes that is compatible with each other agent (who fulfils the normative requirement) will be vindicated in the limit, or even vindicated across all courses of inquiry in the limit.

In a theory of formal learning for the hypotheses of a scientist there are some paradigms or contexts in which the scientist is incapable of identifying a correct hypothesis, even in the limit. The point of formal learning theory is to discriminate between cases in which learning is possible in the limit and ones in which it is not. Thus, in a full development of the theory of learning that I propose here, we would discriminate between cases in which communicative rationality would succeed in identifying norms and needs in the limit from cases in which it would not.

Summary

This paper has considered a different paradigm for resolving disputes, or problems, about meaning from a simple paradigm of negotiation. The difference consist in a special protocol with side-constraints that prohibit agents from simply seeking to follow their own aims. This second paradigm was introduced at the beginning of the paper by comparison with arbitration. Through its side-constraints, the protocol was to provide for a form of collective arbitration in which there was no need for a separate third-party arbitrator. The protocol was to enable agents to resolve disputes or problems by reference to their needs rather than their original aims. Thus aims were treated as malleable. A good arbitrator provides solutions that are ultimately persuasive and rationally so. Thus a solution respects aims and side-constraints that will ultimately be articulated under conditions of rationality rather than the aims which disputants have before their dispute is resolved. An account of agents

was sketched in which the aims of agents were malleable. The account was far from providing a blueprint for constructing such agents, but nevertheless it was suggested that artificial agents might be constructible on refinements or abstractions from it. The account was derived from Habermas' theory of communicative rationality, and a simple version of his position was sketched in the paper. At the end, a sketch of a formal theory of learning for communicative rationality, involving learning of norms and needs, as well as hypotheses, was provided. The idea was that it might be developed into a formal framework for distinguishing cases in which communicative rationality would succeed from cases in which it would not.

References

Austin, J. 1975. *How to do things with words*. OUP, 2nd. edition.

Bratman, M. 1987. *Intentions, Plans and Practical Reason*. Harvard University Press.

Habermas, J. 1979. *Communication and the Evolution of Society*. Heinemann.

Habermas, J. 1998. *On the Pragmatics of Communication*. MIT/Polity Press.

Jain, S., and al. 1999. *Systems that Learn*. MIT.

Kelly, K. 1996. *The Logic of Reliable Inquiry*. OUP.

Nozick, R. 1974. *Anarchy, State and Utopia*. New York: Basic Books.

Young, R. forthcoming. Context in philosophy of science. In Bouquet, P., S. L., ed., *Research in Context*. CSLI.