Communication Breakdown? Reasoning about Language and Rational Agents

Rodger Kibble[1]

Abstract. This paper examines different approaches to rationality in analytic philosophy and AI, in the light of Bermúdez's proposal that a full account of rationality must aim to explain how agents can both select and explain actions, as well as assessing them against some normative standard. We briefly survey instrumental, linguistic and discursive accounts of rationality, and conjecture that Habermas's notion of the "three roots" of epistemic, teleological and communicative rationality comes closest to providing a satisfactory account, or at least the ingredients of such an account. This is contrasted with the widely-accepted BDI model of rational agency in AI, which we argue falls short of a full model of rationality and in particular, fails to provide a convincing model of linguistic communication.

1 Introduction

A major paradigm in AI research has been the development of so-called 'rational', 'autonomous' agents, with a degree of consensus around the BDI architecture [28, 29, 39] which aims to develop software entities that are capable of pro-attitudes such as Beliefs, Desires and Intentions. As Bermúdez notes, the concept of rationality finds applications in various disciplines

> from economics to political science, fron philosophy to psychology, and from management science to sociology [2, p. 1]

while it is not clear whether the word is used with the same sense in all these different contexts. In this paper we will explore how rationality is understood and applied within Artificial Intelligence, against the setting of a brief survey of accounts of rationality from analytic philosophy.

The ability to conduct fluent natural language conversation has been considered a hallmark of human-level AI (or AGI) since Turing's classic paper "Computing Machinery and Intelligence" [37], while there is a long-standing view in some schools of philosophy that thinking, rational agents must be able to express their thoughts in language and interpret the speech and thoughts of other agents. One strand of research in computational linguistics (CL) builds on Searlean Speech Act Theory (SAT), and formalisations of SAT as a CL programme fed in turn into development of communication languages for artificial software agents (Agent Communication Languages or ACLs) [26].

1.1 Outline of the paper

We begin in Section 2 by sketching some different conceptions of what is meant by rational agents in AI and philosophy, discussing instrumental, linguistic and discursive notions of rationality. Section 3 very briefly summarises some essential characteristics of Anglo-American Speech Act Theory, and Section 4 deals with some classic formalisations of SAT. Section 5 shows how some of this work fed into the specification of communication languages for artificial agents. Throughout the paper we note a tension between agent autonomy and the requirements of successful communication, and conclude in section 6 that this tension is never fully resolved.

2 What are Rational Agents?

As noted above, a significant strand of AI research over the last couple of decades has concerned the development of autonomous agents, which may consist of software-controlled robots or of "disembodied" software entities [28, 39]. Agents are said to be autonomous and rational to the extent that they act independently of specific instructions from their human operators or "principals" and do not rely on extensive built-in knowledge, but make their own decisions based on logical and/or mathematical reasoning using information derived from their perceptions of the state of the world [29]. The term "rational agents" has been used somewhat loosely, and this section looks at some differing ideas of what it means for a system, agent or person to be considered rational.

2.1 Desiderata for rational agents

Bermúdez (op. cit.) questions whether there is a unitary notion of rationality, and proposes that any satisfactory account must serve "three different explanatory projects":

- The project of guiding action: given an agent's information about the state of the world, the actions which can be performed and the expected consequences of these actions in the given conditions, a theory of rationality should idenify a "privileged subset" of actions which will bring about the preferred outcome. This is the notion of rationality which, according to Bermúdez, most interests decision theorists.
- The project of normatively assessing action: what falls under assessment is not only the agent's choice of actions, but the quality of their information-gathering, probability judgments, whether they have succumbed to "framing effects" and so on. Many philosophers think of rationality in these terms; some would interpret "normativity" more broadly to encompass social norms of what is considered to be appropriate behavious (e.g. Brandom [3, 4]).
- The project of explaining/predicting action: this involves either working back from an agent's observed actions to a

[1] Goldsmiths University of London, UK, email: r.kibble@gold.ac.uk

characterisation of their beliefs, desires and other psychological states, or working out what an agent is likely to do given a knowledge of their psychological profile [9]. This is stated to be the aspect of rationality of most interest to psychologists and economists.

The question poses itself: which conception of rationality is most relevant to AI? We might well answer: all of them. Bermúdez in fact concludes that the three dimensions are interpendent, and a satisfactory theory must address them all. However, he argues persuasively that decision theory cannot serve the three explanatory projects simultaneously, and so cannot form the basis of a unitary theory of rationality. In what follows we look at various existing accounts of rationality, and consider whether they can be carved up in a way that corresponds to these projects.

2.2 Instrumental rationality

A leading AI textbook [29] defines a "rational agent" as an entity that perceives and acts in an environment, and acts so as to maximise the expected value of some performance measure. There are various levels of agency:

1. A *reflex agent* responds directly to percepts according to a series of if-then rules which could be encoded as a lookup table, without planning ahead or modelling any aspect of the environment other than its immediate percepts.
2. A *model-based reflex agent* keeps track of the state of the world via internal representations.
3. A *goal-based agent* acts in order to achieve certain goals. An example would be a self-driving car, which has the goal of transporting people or goods from one location to another: the agent would need a repertoire of actions and a knowledge base of how to respond to certain types of percepts, in order to reach its destination safely without contravening any rules of the road or causing harm to its occupants or other persons.
4. A *utility-based agent* tries to maximise some measure of utility or "happiness": for instance if there are different ways of achieving the same goal, a utility-based agent may aim to select the method which involves the least effort or cost. For example, a self-driving car might present its user with different options for a performance measure: taking the fastest route, or one which is predicted to consume the least fuel, the route with the most agreeable scenery, and so on.

This "egocentric" concept of rationality is known as *instrumental*, *goal-oriented* or *strategic* and is generally accepted within "classic" AI, rational choice theory and certain schools of social science [24]. This approach is clearly tailored to the **action-guiding** dimension; though within AI the action-guiding and explanatory/predictive dimensions may sometimes collapse if we assume that actions can be simulated within an exact virtual model of the physical environment, and the agent architecture includes notions of "desire" and "belief".

2.3 Linguistic rationality

It has long been argued that only language-using agents can be capable of reasoning and abstract thought[2]. One proponent of this view is Donald Davidson [11], who maintains that a "rational animal" is one that has propositional attitudes such as belief, desire and intention. His argument is that in order to hold even one propositional attitude, an agent must possess a substantial body of beliefs and conceptual knowledge, and be capable of making logical inferences using these concepts. Only a language-using agent can show evidence of this kind of knowledge, and so only this class of agents can be acknowledged as rational. This line of argument extends to actions:

> Intentional action is action that can be explained in terms of beliefs and desires whose propositional contents rationalize the action. [11, p. 99]

It is not immediately clear whether a utility-based agent as desribed in section 2.2 could be extended with these capabilities without radically redesigning the agent architecture. As Davidson points out, agents such as unmanned autonomous drones behave in the way they do not because they have personal desires or intentions to attack a particular target, but because they are designed and operated by humans who have these desires[3]. He argues in his earlier [10] that this "intentional" interpretation of rationality is more appropriate for action explanations:

> . . . I think that compared to attribution of of desires, preferences or beliefs, the axioms of decision theory lend little empirical force to explanations of action (p. 273)

So we may argue that linguistic rationality is a minimal requirement for the project of **explaining/predicting** actions, if not sufficient in itself. Davidson concludes that rationality is essentially a social trait: "only communicators have it".

It is worth noting at this point that John Searle [32] would not accept complex verbal behaviour as sufficient evidence of consicousness or intentionality, since he argues that the architecture and physical make-up of digital computers render them incapable of such states. Davidson [11] declines to address the question of "whether all that prevents computers from being rational is their genesis", and we will likewise leave this question to one side for the present.

2.4 Discursive or argumentative rationality

Another approach may be termed *discursive* rationality, and characterises agents that are capable of giving reasons for their actions and asking for reasons from other agents. Two proponents of variants of discursive rationality are Robert Brandom [3, 4, 5] and Jürgen Habermas [17, 18, 19, 20]. Both propose that agents who utter speech acts are taking on certain kinds of *commitments*, including a commitment to provide evidence for the truth or acceptability of one's assertions.

Of course, in many circumstances it would not be reasonable to call someone out on these commitments: as Brandom puts it, sometimes doubts too need to be justified. A reviewer gives the example "it's five o'clock and I'm in a hurry writing a paper" - one can understand they would be annoyed if asked to substantiate this claim. Steinhoff [35] likewise argues that the statement "Yesterday I was at home alone" would be"quite believable" in the absence of any proof. This of course depends on the context: if someone answering my description had been caught on CCTV stealing some rare jazz LPs at the same time, a police officer investigating this larceny might be less inclined to give me the benefit of the doubt. Steinhoff also argues that one can be subjectively certain of a claim without being

[2] For example, in HG Wells' novella *The Island of Doctor Moreau* the eponymous Dr achieves human-level intelligence in animals by operating on their vocal tracts so that they are physically capable of articulating human speech.

[3] A distinction which is known in the literature as *original* versus *derived* intentionality; see e.g. [12].

in a position to justify it to others: "If aliens were to 'beam' me to their planet in another galaxy without anyone else noticing, and then 'beam' me back shortly thereafter, then for me the thesis that aliens exist would be a very well-justified thesis, without my being able to justify it to others." I would argue to the contrary that an episode like this which would be quite outside my previous experience, and which if true would violate my understanding of the laws of physics, would be more reasonably interpreted both by myself and others as a vivid dream or hallucination rather than a "first contact" with extraterrestrial beings unless some robust independent evidence were available.

It can be seen that the specifications for goal- and utility-based agents do not include this ability to reflect on and reason about the speech acts of oneself or others. Davidson does not explicitly address this issue in the cited paper but it is arguable that an agent would need the capacity to engage in argumentation in order to qualify for "linguistic rationality".

Brandom's approach is concerned with "deontic" attitudes of hearers, and of speakers as self-monitors, rather than intentional attitudes of speakers as in classic Speech Act theory. In place of beliefs and desires, Brandom discusses "doxastic" (propositional) and practical commitments, which interacting agents may acknowledge or ascribe to one another.

The normative dimensions of language use according to Brandom comprise *responsibility* - if I make a claim, I am obliged to back it up with appropriate evidence, argumentation and so on - and *authority* - by making a claim to which I am assumed to be entitled, I license others to make the same claim. The essential idea is that making an assertion is taking on a *commitment* to defend that assertion if challenged. There are obvious shared concerns with the notions of commitment developed by [21, 38]. Brandom's elaborations include the notion of *entitlement* to commitments by virtue of evidence, argumentation etc; the interpersonal inheritance of commitments and entitlements, and the treatment of consequential commitments and incompatibility. Brandom has a similarly "holistic" conception of semantics to Davidson, maintaining that in order to hold one belief it is necessary to hold several, and to be capable of articulating inferential relations among beliefs.

The mechanism for keeping track of agents' commitments and entitlements consists of *deontic scoreboards* maintained by each interlocutor, which record the set of commitments and entitlements which agents claim, acknowledge and attribute to one another (claims and acknowledgements are forms of self-attribution). Scoreboards are perspectival and may include both explicitly claimed commitments and consequential commitments derived by inference. Thus an agent may be assessed by others as being committed to propositions which are entailed by his overt commitments, whether or not he acknowledges such commitments. In principle, Brandom's framework does not necessitate imputing mental states to agents, as deontic statuses are constituted by agents' public or implicit attributions of commitments and entitlements to one another, and he states at one point "I do not officially believe in belief" [4]. In practice however, he regularly slips into mentalistic talk, particularly when discussing intentions [3, Ch. 3].

According to Habermas, interlocutors in a dialogue implicitly make *validity claims* pertaining to each utterance:

- A claim that the propositional content of an assertion is *true* (*wahr*) or that in the case of non-declarative utterances, the existential presuppositions of the propositional content are satisfied;
- A claim that the speaker is *truthful* (*wahrhaftig*) or reliable;
- A claim that the utterance is *appropriate* (*richtig*) according to any applicable social norms.

What makes an agent rational is the ability and disposition to respond appropriately to challenges to validity claims, by offering various forms of argumentation and evidence, with the aim of reaching mutual understanding or rational consensus. Discursive rationality presupposes the validity of a *normative background* of

> institutions, roles, socioculturally habitualized forms of life – that is, of conventions [18, p. 76]

Social norms of this kind are resistant to formalisation and, I would argue, cannot be reduced to considerations of instrumental reasoning [25]. If we accept the idea that speech acts raise these kinds of validity claims, it would seem reasonable to extend this to actions: agents would be taken as implicitly claiming that their actions are appropriate and effective. Brandom goes some way towards a unitary account of propositional and practical commitments: to perform an action is, at least implicitly, to claim an entitlement to being committed to that action, and may also count as redeeming a commitment and/or taking on further commitments. To take a banal example, if you park your car by the side of the road you are likely to be thereby committed to paying a fee to the local council; if you make appropriate signalswhile driving you are committed to manoeuvring your vehicle in various ways, and so on. In Habermas's framework, we could perhaps construe a rational, intentional action as raising claims to be effective, deliberate and appropriate to prevailing social norms, corresponding to the discursive claims of truth, truthfulness and "rightness".

2.5 Habermas's Three Roots of Rationality

Habermas in his earlier work [17, 18] makes a distinction between strategic and communicative uses of speech acts: the former is intended by the speaker to achieve some end through its perlocutionary effects, which need not be transparent to the hearer, while the latter is intended to reach agreement or understanding with hearers and involves a readiness to vindicate any validity claims as well as to take seriously the claims of others. This classification can be argued to have its roots in ethical considerations as well as communicative efficiency. Kant's practical imperative enjoins us to treat ourselves and other people as ends in themselves rather than means to an end [23], while strategic action can involve treating others as means. And the requirement to justify validity claims when challenged increases the likelihood of arriving at a rational consensus [17].

One could take this further and argue that the possibility of achieving consensus assumes a reciprocal *recognition* of interlocutors as *persons* with equal rights to offer and challenge validity claims. Linguistic philosophers tend to abstract away from such considerations: Searle stipulates as one of the preconditions for successful execution of speech acts that "normal input-output conditions obtain", meaning that the interlocutors speak the same language and can hear each other clearly; neither is physically impaired or under duress, acting in a play, telling a joke and so on. Brandom's framework of commitments and entitlements appears to tacitly assume that participants in discourse accord each other the same rights and opportunities to be heard and to raise questions.

In later work [20], Habermas recognises three "roots of rationality" which in some ways reflect the distinctions above between instrumental, linguistic and discursive rationality, as well as Bermúdez's projects of selecting, explaining/predicting and assessing actions:

- **Epistemic rationality** concerns propositional or procedural knowledge of facts about the world and ways of doing things, which informs rational, goal-directed actions. This seems to approximately correspond to Davidsonian linguistic rationality, and is arguably a requirement for the project of explaining/predicting actions.
- **Teleological rationality** concerns the deliberate and purposeful selection and implementations of means to achieve a desired result, on the basis of particular preferences. This is somewhat similar to what we have labelled as instrumental rationality, except that Habermas claims that this deliberation necessarily rests on linguistically structured knowledge, intentions and inferences.
- **Communicative rationality** is essentially what we have called "discursive rationality": the ability and disposition to raise, challenge and defend validity claims of truth, truthfulness and appropriateness. This could be argued to be a prerequisite for the normative assessment of actions.

These three roots are conceptually distinct but are claimed to be equally fundamental to the constitution of rational agents. At some risk of forcing the similarities, one could conjecture that this three-pronged approach equips us to tackle Bermúdez's three explanatory projects - though of course this remains to be worked out in detail. By contrast, the BDI framework aims to build linguistic agents "on top" of utility-oriented, goal-seeking agents or planning systems. Within the confines of this paper, we will look at one line of development in languages for artificial agents and conclude that it does not succeed in modelling successful communication.

2.6 The BDI framework

The BDI framework [28, 39] seems to combine elements of Habermas's epistemic and teleological rationality, in that agents are said to be rational if they choose actions which are

> in [their] own best interests, given the beliefs [they have] about the world. [39, p. 1]

The framework claims to model agents with "mental states" such as beliefs, desires and intentions, but it is not clear that these entities possess or manipulate propositional or conceptual knowledge as stipulated by Davidson and Brandom as a prerequisite for rationality. Rather, "beliefs" in these systems are data structures representing some aspects of the state of the world, while "desires" and "intentions" represent preferences and a currently chosen course of action. Agents can be considered "rational" according to the BDI model without necessarily having linguistic capabilities: these are typically implemented by treating language itself as a kind of action, following variants of Austin and Searle's Speech Act theories as discussed below. It is generally assumed in these models that intentions are constrained by beliefs and desires: that is, an agent's intended or planned actions will be governed by what it believes the current state of the world to be, how it desires the state of the world to change and what actions it believes can be successfully executed. It might be more psychologically plausible to consider in addition whether beliefs can be moulded by desires and intentions, particularly when agents have powerful desires and their beliefs are based on partial or inconsistent evidence: for example a political leader wishing to overthrow a tyrant might be more disposed to give credence to claims that the tyrant possesses fearsome weapons which can be readily mobilised [33].

3 Communicating Agents

This section looks at some applications of Speech Act Theory to the specification of communication languages and protocols for intelligent software agents (ACLs) originating with Cohen and Perrault's seminal 1979 paper [8]. Some desiderata for agent communications were listed by Cohen and Levesque [7]:

> A language for interagent communication should allow agents to enlist the support of others to achieve goals, to commit to the performance of actions for other agents, to monitor their execution, to report progress, success and failure, to refuse task allocations, to acknowledge receipt of messages, etc.

Cohen and Perrault proposed that the essential characteristics of speech acts as described by Searle [30] can be captured using classic AI planning operators and knowledge representation, techniques which fall under the headings of goal- and utility-oriented agents as outlined above. The question thus arises whether linguistic and discursively rational agents can be developed within an architecture for instrumental agency. We discuss below whether work in computational linguistics and agent design following on from Cohen and Perrault has succeeded in in modelling successful communication among autonomous agents, and we conclude that no convincing solution has been found. Firstly, in the next section we briefly outline some essentials of Speech Act theory.

3.1 Speech Acts and Communicative Action

The idea that utterances or "locutions" can have the effect of *actions* which change the state of the world goes back at least to Hegel's Lectures on the Philosophy of History [22]:

> Reden aber sind Handlungen unter Menschen und zwar sehr wesentlich wirksame Handlungen.
> Speeches, it must be allowed, are veritable transactions in the human commonwealth; in fact, very gravely influential transactions.

Hegel here considers the impact of speech on the world-historical level; the 20th-century Anglo-American programme of Speech Act Theory, which developed out of Oxonian "ordinary language" philosophy, showed that it could be fruitful to view utterances as actions on a more parochial level. A special case of speech acts or illocutionary acts is the *performative*, where pronouncing a sentence has the effect of performing the action it appears to describe or refer to:

1. I pronounce you man and wife.
2. I promise I will come to the seminar.

Whether such acts are performed successfully or appropriately depends on factors such as the speaker's social role, their sincerity and so on: for instance only an ordained minister or authorised government employee can correctly utter (1). Other types of speech acts lack the performative feature but may still be seen as actions which change the world in some way: for example, questions or requests can have the effect of imposing some obligation on the hearer, when uttered in appropriate circumstances:

3. What is the capital of Peru?
4. Please come to the seminar.

3.2 "Classic" Anglo-American Speech Act Theory

The essence of Speech Act Theory is the assumption that agents as speakers (S) typically produce utterances with the intention of bringing about some change in the beliefs of a hearer (H), and that Hs recognition of this intention is crucial to the success of the speech act. Speech acts are sub-categorised by their **preparatory conditions** and **essential conditions**; for instance the preparatory conditions for assertion of a proposition p are that S has evidence for the truth of p and it is not obvious that H knows p, while the essential condition is that uttering an assertion counts as an undertaking to the effect that p represents an actual state of affairs [30]. For Habermas [18] the illocutionary force serves to establish a social relation between interlocutors: for instance a request to borrow money seeks to set up a relation between creditor and debtor.

Speech act theory since [1] has distinguished between illocutionary effects including changes in the hearers beliefs which are conventional consequences of an utterance, and perlocutionary effects including consequences of an act which may or may not have been intended by the speaker. According to Searle [30], utterances can be broken down into their *propositional content* and their *illocutionary force*, which classifies acts as assertions, requests, questions, commands and so on.

3.3 Agents and Agent Communication

We shall see in the following survey that there is a tension between **agent autonomy** and the requirements of successful communication. Wooldridge [39] points out a paradox in mentalistic speech act based semantics for the inform locution in communication between autonomous software agents:

> If I am *completely* autonomous, and exercise complete control over my mental state, then nothing you say will have any effect on my mental state (...) if you are attempting to inform me of some state of affairs, then the best you can do is convince me that you believe this state of affairs.

As observed by [26], it is not clear that this roundabout approach succeeds in resolving the "paradox of communication". If the hearer agent is "completely autonomous", then convincing it that you believe P or even that you intend it to believe P seem just as problematic as getting it to believe P itself: both of these are still an attempt by the speaker to bring about a change in the hearer's mental state.

4 Initial formalisations of Speech Act Theory

Searle's [30] formulation of necessary and sufficient conditions for illocutionary acts is certainly an advance over Austin's [1] original proposals in its relatively explicit and systematic nature. Indeed, Searle with Daniel Vanderveken [31] developed a formalization of the theory known as Illocutionary Logic. However, subsequent implementations of Speech Act Theory in computational dialogue modelling and intelligent agent communication did not develop directly from this work, but rather from work rooted in the AI tradition of planning, reasoning and knowledge representation. Cohen and Perrault [8] showed in a now classic paper that there is a compelling similarity between Searle's necessary and sufficient conditions for performance of speech acts and the pre-and post-conditions used to define planning operators in AI. Cohen and Perrault were explicitly concerned to develop the beginnings of a "competence theory of speech act communication" by presenting planning operators and inferential rules for plan construction that "should lead to the generation of plans for those speech acts that a person could issue appropriately under the same circumstances".

In order to be able to model the successful performance and uptake of selected speech acts, they adopt a simplifying assumption of complete cooperativity among interlocutors, such that for example to request someone to do something is sufficient to get them to want to do it.

In Cohen and Perrault's system the preconditions for planning operators capture Searle's preparatory and sincerity conditions, while the effect models the essential condition. The authors state that speakers "cannot influence their hearers beliefs and goals directly", and the INFORM act is accordingly defined so that it can be performed successfully without the hearer necessarily actually coming to believe the asserted proposition. The CONVINCE act is defined to demonstrate that perlocutionary effects can be achieved if we assume that agents have complete trust in informants' truthfulness and reliability. There is a snag here, however. The definition of INFORM still specifies that the hearer's mental state is updated as a result of the action, with the belief that the speaker believes what they say. This runs contrary to the principle that other agents' world models cannot be directly modified by communicative actions.

> While a speaker often has performed illocutionary acts with the goal of achieving certain perlocutionary effects, the actual securing of those effects is beyond his control (Op cit: 187)

It turns out that in a multi agent environment, the securing of illocutionary effects on the recipient may also be beyond the control of a communicating agent. Thus we may question whether the distinction between illocutionary and perlocutionary acts is a useful one in this context.

Cohen and Levesque [6] have the ostensible aim of presenting a logical theory which "predicts dialogue phenomena" and can be applied to the development of "algorithms for human-computer interaction in natural language". In fact the theory is presented at some level of abstraction from any natural language constructions: the paper contains very few natural language examples, mostly simple imperatives such as "Open the door".

Note that C&L treat all communicative acts as "attempts", i.e. nothing in the definition of an act entails any changes in the addressee's goals or intentions, which would be incompatible with the addressee's status as an autonomous agent. Success of an illocutionary act requires not only that the utterer has performed the act correctly, but in addition that interlocutors can be assumed to be sincere and helpful; formal definitions of both of these characteristics are provided.

In summary, Cohen and Levesque's system does not seem to be vulnerable to the objections we have noted as applying to Cohen and Perrault [8] since they do not define actions which directly modify the state of another agent; rather, the success or failure of an action depends on the addressee's level of helpfulness and the extent to which the addressee regards the hearer as sincere. Interestingly, the authors define two degrees of success criteria which seem to match the distinction between illocutionary and perlocutionary effects: a requester is committed to his request being understood, in that he will be likely to repeat it if it seems that the addressee has not heard him clearly; but it is not required that the agent is committed to the perlocutionary goal of getting the addressee to perform the required action. Essentially, it appears that an illocutionary act may be accepted or rejected by a recipient, while the sender has the option of repeating the same action or giving up. There appears to be no mechanism

for receivers to challenge or query on action on specific grounds, or for the sender to present an argument or further evidence to persuade the receiver to accept a request.

5 Communication languages for software agents

The most ambitious programme for developing an artifical communication regime along the lines of SAT was the FIPA ACL [16, 15, 14, 13]. Communicative acts are specified in terms of their **feasibility preconditions (FP)** which correspond to Searle's preparatory and sincerity conditions, and **rational effect (RE)** specifying the intended outcome of a CA. The latter corresponds to the perlocutionary effect: for example the RE of an assertive act is that the hearer comes to believe the asserted proposition, not simply to be aware that the speaker believes it. Since perlocutionary effects are not under the direct control of the sender, REs are generally not regarded as deterministic but are qualified with statements like the following [14]:

> Whether or not the receiver does, indeed, adopt belief in the proposition will be a function of the receivers trust in the sincerity and reliability of the sender.

This in fact seems rather too loosely worded, as surely the receiver would also take into account any independent evidence it has as to the truth of the proposition: the sender may be utterly sincere and normally reliable, but simply mistaken in a given instance. And given that the sender may not be aware how much it is trusted by the receiver, it is not clear how it should decide whether to update its model of the receivers beliefs – in any case this is not covered in the specifications. Illocutionary effects are not specified for individual communicative acts but are partially implemented via an over-arching axiom called Property 4: "When an agent observes a CA, it should believe that the agent performing the act has the intention (to make public its intention) to achieve the rational effect of the act. This is called the 'intentional effect'" [14]. To spell things out: an agent i should believe that if agent j has performed an action a of a particular type,j intends that the defined rational effect of this type of action should come about.

The FIPA specifications envisage that agents will have the capacity to reason about other agents' knowledge and beliefs; for example, an agent i seeking to inform agent j of proposition p should choose one of three different communicative acts, depending on its model of j's prior beliefs or lack of belief concerning p.

How can agents reason about each others' beliefs? Some sources of evidence are utterances that agents have produced, and utterances they have received. In the first case, an agent's beliefs should match the preconditions for any communicative acts that it executes, and these beliefs are assumed to persist after the act has been performed according to an axiom designated "Property 5" [14]. In the latter case, the agent's beliefs may match the REs of communications they have received, though this would count as weaker evidence.

Preconditions which require the sender to have knowledge of the receivers state of mind are not realistic as the sender cannot verify whether they actually hold; nor, a fortiori, can the sender directly determine whether the RE has been achieved. It turns out however that the full specifications do not assume that agents' behaviour is based solely on semantic reasoning; rather a number of predefined protocols are available which mandate response to particular message types [13]:

> The protocol parameter defines the interaction protocol in which the ACL message is generated. This parameter is optional; however, developers are advised that employing ACL without the framework of an interaction protocol (and thus directly using the ACL semantics to control the agents generation and interpretation of ACL messages) is an extremely ambitious undertaking.

The FIPA ACL specification includes interaction protocols for certain common sequences of message exchanges, enabling agents to cut down on the extent of reasoning required at any step: rather than forward-chaining from the current state of a model to find appropriate actions to achieve their goals, they can apply backward-chaining from a restricted set of available options specified by the protocol in order to find which, if any, has FPs that match the current state.

A running theme of this paper has been the tension between agent autonomy and the requirements of successful communication, which manifests itself in several ways: genuinely autonomous agents can never be certain that an interlocutor's mental state matches the preconditions for an utterance, nor that the required effects have been achieved. As Poslad [27] observes, the specifications for communicative acts in FIPA ACL require the sender to respect the FPs in order to send a message, but do not oblige the receiver to respect the RE:

> The actual interpretation of the sender's intentional effect in the receiving agent is considered to be relative to each agent and to be customizable by each agent. Algorithms for BDI rule engines to interpret the sender's intent at the receiver are not specified by FIPA, although, this has been proposed and discussed several times at FIPA meetings. The semantics is underspecified in the sense that whilst receiving agents receive CAs concerning the intentions and beliefs of the sender, receiving agents are free to carry out their internal actions, such as changing beliefs, which may be consistent or inconsistent with the sender's CA.

We have seen variants of assertive actions which seem to fall between two stools: on the one hand they aim only at causing the receiver to believe that the sender believes the asserted proposition, rather than getting the receiver to believe the proposition itself; on the other, even this modest goal is actually too strong if we take agent autonomy seriously. In fact it is doubtful whether BDI agents are actually implementable in their "pure" form [27]:

> BDI models have incomplete axiomizations and can be computationally complex or even intractable. The BDI model focuses on private belief and intention transfer between individuals. It doesnt take into account third party or societal interaction and associated constraints. BDI models seldom focus on pragmatic issues such as belief and intention management. These can make the model computationally complex or even intractable.

Finally we should briefly mention Singh's [34] social semantics for agent communication languages, which is explicitly designed to avoid having to attribute mental states to artificial software agents and is not explicitly intended to model human communication. Singh claims to follow Habermas in distinguishing three levels of validity claim that are raised with each communication: objective (the claim is true), subjective (the sender is sincere, and believes the content of the communication) and practical (the sender is justified in sending the communication, e.g. in taking on a particular commitment, or making a request or prohibition). The cited paper includes formalisations of communicative acts such as inform, request, permit,

forbid, promise and declare. However, although these formalisations include specifications of the agents' commitments or validity claims, there seems to be no locution or protocol for challenging or justifying these claims, so the model cannot claim to faithfully reflect Habermas's original notion of communicative rationality.

6 Conclusion and Future Work

This paper has looked at various ideas of what makes up a rational agent, and we have discussed some evidence that rationality requires three separate, if linked, capabilities:

1. Ability to select actions in order to achieve some aim, whether by maximising utility or formal practical reasoning;
2. Ability to explain and predict one's own actions or those of other agents;
3. Ability to assess actions against some normative standards.

We have conjectured that this distinction roughly corresponds to the classification of rationality models as instrumental, linguistic and discursive, and to Habermas's "three roots of rationality". Habermas is the only one of the researchers we have considered who attempts to pull together different dimensions of rationality within a unitary account, though further work will be needed to establish whether his approach can adequately tackle these three explanatory projects. Finally, we have looked at one particular strand of work in AI which aims to construct rational agents, and argued that the resulting artefacts fall short of being fully rational according to the above criteria.

David Traum's observation still holds true [36]:

> While a comprehensive theory of speech acts will strain most contemporary theories of rational agency, requiring a very expressive theory of action and mental state, it also provides a good testbed for a theory of agency in a multi-agent world.

In fact it is doubtful whether any ACL specification can so far be said to incorporate a "comprehensive theory of speech acts", as many questions have been left unresolved. It has proved chellenging to specify exactly how the intended outcomes of communicative acts should be specified in ways that take full account of the receivers autonomy. The systems we have considered tend to assume an egocentric model of goal-directed agency, in contrast to Habermas and Brandom's conceptions of rational agency which postulate a normative background for communicative action. Each of the implementations and formalisations we have considered turns out to rely on some external mechanism to secure uptake, such as norms of sincerity and cooperativity or mandated communications protocols.

To return to where we started: there is an interesting connection between the notion of discursive rationality and Turing's [37] proposal that the ability to conduct unrestricted conversation in natural language could be taken as indicative of intelligence. This ability would have to encompass giving reasons or justifications for claims as well as demanding them. This is precisely what is missing from the various "chatbots" which have been entered for the annual Loebner Prize: see for example this exchange involving "Mitsuku", the most successful entrant in the 2016 competition[4].

> Judge: The cat tried to climb in the box but got stuck because it was too big. What was too big?
> Mitsuku: Well at least they tried their best. Large, above average in size or number or quantity or magnitude or extent.

[4] See http://www.aisb.org.uk/events/loebner-prize#Results16

The reader will recall that Turing offered few substantive arguments in favour of his proposal, focussing instead on rebutting a series of anticipated objections. If the claims of this paper are accepted, discursive abilities can be taken as a necessary if not sufficient condition for agents to be considered to be rational, adding a positive argument for Turing's proposal alongside his negative ones.

ACKNOWLEDGEMENTS

I would like to thank the anonymous referees for their comments, which I hope have helped to improve this paper.

REFERENCES

[1] J. Austin, *How to Do Things with Words*, Oxford University Press, 1962.
[2] L. Bermúdez. J. *Decision Theory and Rationality*, Oxford University Press, 2009.
[3] R. Brandom, *Making It Explicit: Reasoning, Representing, and Discursive Commitment*, Harvard University Press, Cambridge, MA, 1994.
[4] R. Brandom, *Articulating Reasons: An Introduction to Inferentialism*, Harvard University Press, Cambridge, MA, 2000.
[5] R. Brandom, *Between Saying and Doing: Towards an Analytic Pragmatism*, Oxford University Press, Oxford, 2008.
[6] Philip Cohen and Hector Levesque, 'Rational Interaction as the Basis for Communication', (1988). Technical Note 433, CLSI, Stanford University.
[7] Philip Cohen and Hector Levesque, 'Communicative actions for artificial agents', in *Proceedings of ICMAS-95*, p. 65 72, (1995).
[8] Philip Cohen and Raymond Perrault, 'Elements of a plan-based theory of speech acts', *Cognitive Science*, 177–212, (1979).
[9] Donald Davidson, 'Actions, reasons and causes', in *Essays on Actions and Events*, ed., Donald Davidson, 3–19, (1963/2001). First published in *Journal of Philosophy* 60 (1963).
[10] Donald Davidson, 'Hempel on explaining action', in *Essays on Actions and Events*, ed., Donald Davidson, 95–105, (1982/2001). First published in *Erkenntnis* 10 (1976), pp. 239-253.
[11] Donald Davidson, 'Rational animals', in *Subjective, Intersubjective, Objective*, ed., Donald Davidson, 95–105, (1982/2001). First published in *Dialectica* 36, 1982, pp 317-273.
[12] Daniel C. Dennett, *The Intentional Stance*, MIT Press, 1989.
[13] FIPA, 'ACL message structure specification: Sc00061g, dec. 2002', Technical report, (2002).
[14] FIPA, 'Communicative act library specification: Sc00037j, dec. 2002', Technical report, (2002).
[15] FIPA, 'Query interaction protocol specification: Sc00027h, dec. 2002.', Technical report, (2002).
[16] FIPA, 'SL content language specification: Sc00008i, dec. 2002', Technical report, (2002).
[17] Jürgen Habermas, 'Wahrheitstheorien', in *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns*, 127–186, (1972/1984).
[18] Jürgen Habermas, 'What is universal pragmatics?', in *On the Pragmatics of Communication*, ed., Maeve Cooke, 21–104, (1976/1999). Translated by Maeve Cooke.
[19] Jürgen Habermas, 'Social Action, Purposive Activity, and Communication', in *On the Pragmatics of Communication*, ed., Maeve Cooke, 105–182, (1981/1999). Translated by Maeve Cooke.
[20] Jürgen Habermas, 'Some further clarifications of the concept of communicative rationality', in *On the Pragmatics of Communication*, ed., Maeve Cooke, 307–342, (1996/1999). Translated by Maeve Cooke.
[21] C Hamblin, *Fallacies*, Methuen, London, 1970.
[22] G. W. F. Hegel, *Vorlesungen über die Philosophie der Geschichte*, Reclam, 1989. First published in 1837.
[23] Immanuel Kant, 'Groundwork of the metaphysics of morals', in *The Moral Law*, ed., Paton H. J. Routledge Classics, 53–148, (1948/2005). Translated by H. J. Paton.
[24] Rodger Kibble, 'Homo practicus: Practice theory and post-cognitivist AI', in *Proceedings of AISB Convention, 2016*.
[25] Rodger Kibble, 'Regulism, regularism and some limitations of agent-based modelling', in *Proceedings of AISB 2013*, (2013).

[26] Rodger Kibble, 'Speech act theory and intelligent software agents', in *Pragmatics of Speech Actions*, eds., M Sbisa and K Turner, Walter de Gruyter, (2013).
[27] Stefan Poslad, 'Specifying protocols for multi-agent systems interaction', *ACM Trans. Autonom. Adapt. Syst. 2, 4, Article 15 (November 2007)*, (2007).
[28] Anand S. Rao and Michael P. Georgeff, 'BDI Agents: From theory to practice', in *IN PROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON MULTI-AGENT SYSTEMS (ICMAS-95*, pp. 312–319, (1995).
[29] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Pearson International, 2010.
[30] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, Cambridge, London, 1969.
[31] John Searle and Daniel Vanderveken, *Foundations of Illocutionary Logic*, CUP, Cambridge, 1985.
[32] John R. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**, 417–424, (1980).
[33] Ian Sinclair, *The March That Shook Blair*, Peace News, 2013.
[34] M.P. Singh, 'A social semantics for agent communication languages', in *Issues in Agent Communication*, pp. 31–45, (2000).
[35] Uwe Steinhoff, *The Philosophy of Jürgen Habermas: A Critical Introduction*, Oxford University Press, 2009.
[36] David Traum, 'Speech acts for dialogue agents', in *Foundations And Theories Of Rational Agents*, 169–201, Kluwer Academic Publishers, (1999).
[37] Alan Turing, 'Computing Machinery and Intelligence', (1951).
[38] D.N. Walton and E.C.W. Krabbe, *Commitment in dialogue: basic concepts of interpersonal reasoning*, SUNY series in logic and language, State University of New York Press, 1995.
[39] Michael Wooldridge, *Reasoning About Rational Agents*, MIT Press, 2000.