A Cognitive Model for Conversation

Nicholas Asher
CNRS, Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, Toulouse
asher@irit.fr

Alex Lascarides
School of Informatics,
University of Edinburgh
alex@inf.ed.ac.uk

Abstract

This paper describes a symbolic model of rational action and decision making to support analysing dialogue. The model approximates principles of behaviour from game theory, and its proof theory makes Gricean principles of cooperativity derivable when the agents' preferences align.

1 Introduction

Grice (1975) and Neo-Griceans model the link between dialogue processing and general principles of rational behaviour by assuming that agents abide by a strong cooperativity principle—namely, people normally believe what they say and help other agents achieve the goals that they reveal through their utterances. This principle provides cooperativity on at least two levels: a basic level that ensures coordination on the conventions governing linguistic meaning (basic cooperativity); and a level concerning shared attitudes towards what is said, including shared intentions (content cooperativity). But not all conversations are content cooperative. For example, Tomm and Dave don't share intentions in (1), taken from chat recordings of an online version of the game *Settlers of Catan* where players negotiate over restricted resources:

(1) a. Tomm: Got any clay to trade for sheep/wheat?
b. Dave: Only got 1 and I'm holding on to it, sorry.

However, even though *Settlers* is a win-lose game where players' interests are often opposed, its players often do share intentions, as they must cooperate to bargain for resources they need in the game.

(2) a. William: can i get a sheep or a wheat?
b. i have too much wood.
c. Cat: i can give you a wheat.
d. William: good
[they exchange 1 wheat for 1 wood]

Conversely, dialogue (3) is content cooperative (and basic cooperative) on the assumption that A and B are constructing a plan to achieve the same goal—that they both eat at Chop Chop:

(3) a. A: Let's go to Chop Chop by car.
b. B: But there's no parking.
c. A: Then let's take the bus.

But (3b) implicates that B *rejects* the intention underlying (3a)—to go to Chop Chop by car. The grounds for the rejecting moves in (1) vs. (3) are different. In (1) Dave can fulfil Tomm's intention but chooses not to, presumably because of his conflicting preferences. In (3) A and B (transiently) have different intentions because of their conflicting beliefs about the optimal way to achieve a shared preference: to eat at Chop Chop.

There have been several attempts to make models of conversation that abide by Gricean principles of cooperativity formally precise; for instance, by expressing axioms in a logic that supports defeasible reasoning about the cognitive states of dialogue agents (e.g., Schulz (2007), Asher and Lascarides (2003)). Such models include (default) axioms Sincerity (following Grice's maxim of Quality (Grice, 1975, p46)): if

a cooperative agent says something then he normally believes it. They also include axioms at the level of intentions. For instance, the following default axiom of Strong Cooperativity is adapted from Grice's analysis of what an utterance *means* in cooperative conversation (Grice, 1969, p151): if an agent says something that implies he has a particular intention ϕ that he also intends should be recognised, then a cooperative agent should normally adopt that intention ϕ too (e.g., (Grosz and Sidner, 1990, p430), Asher (in press)). But such formalisations are incomplete because they do not handle cases like (1) where content cooperativity breaks down: the default axioms just given don't apply in such contexts. Nor do they predict when adopting shared intentions is optimal in a strategic setting (e.g., (2)). Furthermore, (3) shows that, rather than expressing content cooperative behaviour in terms of the default adoption of the other agents' *intentions*, it would be better to define it in terms of shared *preferences*; that way, rejection is rational when conflicting beliefs yield a different optimal way to achieve a shared preference.

This paper provides a cognitive model within which one can explore reasoning about the mental states of dialogue agents. We will derive Gricean principles of cooperativity, formalised as defeasible principles, from a characterisation of certain games, using game theory as the foundation of strategic reasoning and also as the basis for linking inferences about conversation to rational action. In game theory, agents act so as to *maximise their expected utility—utility* being a measure of preference, and the term *expected* ensuring that decisions about action are made relative to one's beliefs about what the outcomes of the actions will be, including beliefs about what other agents will do. Nevertheless, we argue in Section 2 that game theory on its own provides an incomplete picture, which makes it difficult to use to derive defeasible principles. The symbolic cognitive model presented in Section 3 addresses this problem. It provides axioms approximating rational behaviour from game theory that link dialogue actions and mental states, and its *proof theory* allows us to derive Gricean principles of cooperativity when the agents' preferences align. We relate this approach to prior work in Section 4 and point to future work in Section 5.

2 Our Model

Our model for strategic agents is one that is based on logic and on game theory. Like many others, we use a variant of a Belief Desire Intention (BDI) logic to formalise Gricean implicatures. But because we countenance misdirection and deception as features of strategic conversation, we draw a distinction between **Public and Private** attitudes and thus introduce a new attitude for public commitment. Speaking makes an agent *publicly commit* to some content (Hamblin, 1987). Traditional mentalist models of dialogue, couched within BDI logics, *equate* dialogue interpretation with updating mental states: e.g., interpreting an assertion that p is equivalent to updating one's model of the speaker's mental state to include a belief in p (e.g., Grosz and Sidner (1990)). But they are not equivalent in (4):

(4) a. Loreleil292: Can anyone give me some clay for some wheat?
b. AMI123: Sorry have none of that! *[in fact, she has 2 clay]*

We interpret (4b) as a negative answer *even if* we know AMI121's beliefs are inconsistent with this. To do justice to this, we follow Asher and Lascarides (2003) and separate the representation and logic of dialogue content from that of cognitive states and then link them via defeasible transfer principles. This separation was originally motivated by calculable implicatures being unavailable as antecedents to surface anaphora; insincerity provides a new motivation.

In common with game theoretic models of conversation (e.g., Parikh (2001)), we adopt a second principle: people say things that will maximise their expected utility. So if the Gricean maxims of conversation hold, they do so because they maximise the agents' expected utility. We also maintain that agents' preferences evolve as dialogue proceeds, at least partly because agents learn about other agents' preferences from what they say and then adjust at least

some of their preferences in the light of this information. People's preferences are typically **partial** and get more specific or evolve as they learn more through conversation.

This last assumption poses a problem for orthodox game theory. Game theory assumes each player has a completely defined preference function over the possible actions in the game. It models uncertain and partial information that one player has about another player's preferences and the actions that other play is contemplating performing by a probability distribution over player types, where each type is associated with a complete set of actions and a complete utility function.[1] Game theory, however, does not provide general principles for restricting the set of player types one needs to consider or the probability distributions over them. This gap has bite in modelling conversation, because the possible signals that grammars of natural languages allow are unbounded, as are the coherent signals in context. So dialogue agents generally face the task of isolating their game problem to a set of signals that is small enough to effectively perform inference over, but large enough to yield reliable decisions about optimal actions.

To represent dialogue processing it would be better not to remain silent on how one identifies which player types—and hence which actions and preferences—are relevant, but rather to consider a partial theory or description of the agent's preferences that is updated or revised as one learns more about the agent or one considers actions that one didn't consider before. This is what we do here. This approach yields a more compact and tractable cognitive model and a proof theory in which we can reason, in the light of new evidence, about what type of player to consider in our reasoning. Standard game-theory provides models that can verify the soundness of our proof theoretic reasoning. Whether the sentences in this theory are assigned probabilities is not terribly relevant. But what *is* important is that elements of this theory get revised in the light of new evidence, as Alchourrón et al. (1985) suggest. This can either be done by conditionalising a probability distribution over new evidence, or more symbolically via a theory that incorporates general but defeasible principles about human action and the preferences that underlie them.

In our symbolic model, instead of a probability distribution over every possible complete model of the game we begin with just one partial model. We will describe this model in a way that meshes easily with inferring preferences from observing what agents do. We demonstrate one advantage of this approach here: the proof theory afforded by our symbolic axioms of rational behaviour, which approximate those from game theory, is sufficient to *derive* Gricean principles of cooperativity, among them the default axioms of Sincerity and Strong Cooperativity that we discussed in Section 1. We thus gain a logical link between strategic conversation and content cooperative conversation.

3 Cognitive Modelling

To reason about an agent's motives and actions we use a familiar modal logic: $\mathcal{B}_a\phi$ means agent a believes ϕ, and $\mathcal{I}_a\phi$ means a intends to bring about a state that entails ϕ. We assume that $\mathcal{B}_a$ abides by the modal axioms KD45 (so its accessibility relation in the model is transitive, euclidean and serial); so an agent's beliefs are mutually consistent with one another and closed under logical consequence, and agents have total introspection on their beliefs or lack thereof. We make $\mathcal{I}_a$ abide by the modal axiom D (so its accessibility relation in the model is serial), so contradictory intentions are ruled out. We also assume that intentions are doxastically transparent: i.e., $\mathcal{I}_a\phi \leftrightarrow \mathcal{B}_a\mathcal{I}_a\phi$ is an axiom. We also need a modal operator for public commitment, which is distinct from belief: $P_{a,D}\phi$ means agent a publicly commits to ϕ to the group of agents D. Following Asher and Lascarides (2008), we make $\mathcal{P}_{a,D}$ K45 (one commits to all the consequences of one's commitments and one has total introspection on commitments or lack of them). Unlike belief, commitments can be contradictory because one can declare anything.

[1] Game theory allows players to have imperfect knowledge of what action other players play, but that is not relevant here.

Reasoning about mental states is inherently defeasible, so we add to our logic the weak conditional from Commonsense Entailment (Asher, 1995): $A > B$ means *If A then normally B*. We call this language CL (standing for *cognitive logic*). This logic has many nice properties; for instance, soundness completeness and decidability (Asher, 1995). Decidability is maintained even in a dynamic version of the CL (Asher and Lascarides, 2011), but for the sake of simplicity we will consider the static version here.

The dialogue's logical form creates public commitments in CL: if the logical form stipulates that agent a is committed to K at turn n—so K is the content of a clause or of coherently related dialogue segments—then this makes $\mathcal{P}_{a,d}\mathcal{K}$ true in CL, where $\mathcal{K}$ is the CL-representation of the formula K in the separate logic of dialogue content, and D is the set of dialogue agents (how K maps to $\mathcal{K}$ is detailed in (Lascarides and Asher, 2009) but doesn't concern us here).[2]

3.1 Preferences

Besides a representation of dialogue content and BDI attitudes, we need a symbolic way of representing preferences and commitments to preferences. CP-nets (Boutilier et al., 2004) provide a useful formalism for extracting commitments to preferences from utterances (Cadilhac et al., 2011). Standard CP-nets capture *complete* information: they are a compact representation of a preference order over all the possible outcomes of actions that agents can perform. To represent partial preferences, we build a *partial description of a CP-net* (Cadilhac et al., 2011), which approximates preferences as revealed by dialogue moves. This avoids having to postulate a range of player types, each associated with complete preferences. Instead, agents will reason with and revise partial descriptions of preferences as they observe new evidence through dialogue moves.

A CP-net for an individual agent has two components: a directed *conditional preference graph* (CPG), which defines for each feature F its set of parent features $Pa(F)$ that affect the agent's preferences among the various values of F; and a *conditional preference table* (CPT), which specifies the agent's preferences over F's values for every combination of values in $Pa(F)$ (thus CP-nets have a similar structure to Bayesian belief networks (Pearl, 1988)). The CP-net for a *game* consists of a CP-net for each player. For example, the CP-net in Figure 1 represents a game where A chooses what A and B will eat, and B chooses what they will drink (they must eat and drink the same thing). Agent A's preferred *Food* is fish, but the *Wine* he prefers is dependent on the food: white wine for fish and red for meat. Agent B is indifferent about what he eats, but like A his choice of *Wine* is dependent on what he eats. The logic of CP-nets follows two ranked principles when generating the preference order over every outcome from this compact representation: first, one prefers values that violate as few conditional preferences as possible; and second, violating a (conditional) preference on a parent feature is worse than violating the preference on a daughter feature. So Figure 1 yields the following partial order over all outcomes for each agent:

The CPG

Preferences for A:
fish $\succ_A$ *meat*
fish: *white* $\succ_A$ *red*
meat: *red* $\succ_A$ *white*
Preferences for B:
fish $\sim_B$ *meat*
fish: *white* $\succ_B$ *red*
meat: *red* $\succ_B$ *white*
The CPTs

Figure 1: A CP-net for the food and drink game.

$$
\begin{aligned}
&(\textit{fish}, \textit{white}) \succ_A (\textit{fish}, \textit{red}) \succ_A \\
&\qquad (\textit{meat}, \textit{red}) \succ_A (\textit{meat}, \textit{white}) \\
&\{(\textit{fish}, \textit{white}), (\textit{meat}, \textit{red})\} \succ_B \\
&\qquad \{(\textit{fish}, \textit{red}), (\textit{meat}, \textit{white})\}
\end{aligned} \tag{5}
$$

There are efficient algorithms for identifying the (unique) optimal strategy in this case (e.g., Bonzon (2007)): i.e., to eat fish and drink white wine.

Dialogue interpretation yields commitments to preferences that are *partial*. For example, by

[2] K captures *all* a's current commitments, including ongoing commitments from prior turns. So there is no need to conjoin an agent's commitments from each turn in CL.

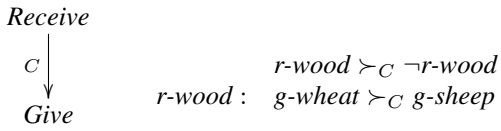

Figure 2: Cat's commitments to preferences in dialogue (2). *Receive* and *Give*'s values are *r-x* and *g-x* where x is *wheat*, *clay*, *sheep*, *rock* or *ore*.

uttering (2c) Cat commits to the partial CP-net in Figure 2: In words, Cat would rather receive wood than not, and given her preference for receiving wood she would rather give wheat than sheep. This is computed recursively from the discourse structure of Cat's commitment (Cadilhac et al., 2011), although we don't detail the mapping here.[3]

Crucially, the CP-net description in Figure 2 is partial: it doesn't reveal Cat's preferences for giving wheat or sheep in a context where she doesn't get wood—(2c) says nothing about that. Cat's actual preferences may also differ from these commitments (e.g., because of insincerity). So to choose an optimal action, agents face a complex calculation in (defeasibly) estimating an agent's complete actual preferences from commitments to them.

Accordingly, we treat the preference statements in the CP-net descriptions as formulas within a background theory that provides defeasible inferences about preferences and behaviour. Our background theory is CL; so CL must be able to express and reason about descriptions of CP-nets. Specifically, we complete the partial information in a CP-net description by adding assumed preferences that defeasibly follow from it via the default axioms in CL, with agents defaulting to being indifferent among values for features for which preference information is missing entirely. In logical terms this means we will have formulae in CL of the form: $\chi > (\phi{:}\,\psi \succ_a \neg\psi)$, where χ is a well-formed formula of CL.[4] In other words, if χ is true then normally the description of a's preferences includes $\phi{:}\,\psi \succ_a \neg\psi$ (note that the antecedent χ may express information about preferences too). We'll give some examples of such formulae in the next section. Further, CL's nonmonotonic inferences about an agent's preferences may change if the range of actions that are considered to be a part of the game changes (though we forego specific examples here). Overall, through the (nonmonotonic) logic of CL's $>$, one can support decisions about what action to perform even if knowledge of preferences is partial.

CL can now link preferences to other propositional attitudes. Indeed, choosing optimal actions requires a link between preference and *belief*: since a (joint) CP-net G can include variables whose values one doesn't control, one needs to check that one's optimal state(s) are not doxastically improbable (this is a crude way of ensuring that agents act so as to maximise *expected* utility rather than acting with wishful thinking about what's feasible). We supply a notion of doxastic improbability in CL via its nonmonotonic consequence relation: i.e., a state is *belief compliant* if its negation does not defeasible follow from the premises and background theory of CL axioms. So to identify an agent's optimal belief-compliant state(s), we filter out any optimal state that is defeasibly inconsistent with his beliefs (as we mentioned in Section 3, this is decidable). Within CL this leads to the definition of a *CP-solution*$_a(\phi, G)$ for agent a and (joint) CP-net G:

Definition 1 CP-solution$_a(\phi, G)$ *holds iff:*

1. *a is a player in the joint CP-net G; and*
2. *$s \vdash \phi$ for every belief-compliant optimal state s of G. I.e., where Γ is the premises—in other words,* CL*'s background theory plus information about the mental states of the players in G—we have $\Gamma \not\vdash \mathcal{B}_a \neg s$, and for*

[3] We take William's and Cat's commitments in (2) to be as follows (Lascarides and Asher, 2009). William's turn commits him to *Plan-Elab*(a, b), which means that he commits to the contents of both (2a) and (2b) and to (2b) elaborating a plan to achieve the goal underlying (2a) (that goal is to obtain a sheep or wheat, and the plan afforded by (2b) is to get one of these by trading wood). Cat's utterance (2c) commits her to *Plan-Elab*(π, c), where π is William's first turn (with content *Plan-Elab*(a, b)). Cadilhac et al's (2011) recursive algorithm yields Figure 2 from *Plan-Elab*(π, c).

[4] Since the features in our CP-nets all take finite values, they can be represented in CL using Boolean variables.

any state s' that is strictly more optimal in G than s, $\Gamma \mid\!\sim \mathcal{B}_a \neg s'$ holds.

For example, if B's model of A's and his own preferences are those in Figure 1, then by Definition 1 *CP-solution*$_B$(*fish*$\wedge$*white*, G) holds: while *meat* $\wedge$ *red* is equally preferred by B, it is not belief compliant because G defeasibily entails that A will choose *white* and not *red*. We'll see this in the next section, when we use CP-solutions to define CL axioms that approximate principles of rational action from game theory.

3.2 Axioms of Rationality

To encode means-end reasoning of rational agents in our symbolic model, we need CL axioms that make agents *pay-off maximisers* (cf. rationality from game theory) and *basic cooperative*. Pay-off maximisers intend actions that are an optimal trade-off between their preferences and their beliefs about what's possible; and an agent intending ψ means in the context of his current beliefs he prefers ψ to all alternative actions. We capture these two principles with the axioms **Maximising Utility** (a) and (b):

Maximising Utility:

a. $(G \wedge \textit{CP-solution}_a(\psi, G)) > \mathcal{I}_a\psi$
b. $(\mathcal{I}_a\psi \wedge \textit{player}(i, G)) > \textit{CP-solution}_a(\psi, G)$

Maximising Utility part (a) ensures a intends ψ only if ψ follows from all belief-compliant optimal states (by Definition 1). Indeed, agent a's intentions are conditional on *all* of a's beliefs (thanks to Definition 1) and *all* of a's preferences and those of any player that affect a's preferences. The latter property follows because the weak conditional $>$ validates the Penguin Principle—i.e., default consequences of rules with more specific antecedents override conflicting defaults from less specific antecedents. So if a more specific game G' is known to hold and it yields conflicting intentions to those resulting from G, then the intentions from G' are inferred and those from G aren't. Axiom (b) likewise conditions a's preference for ψ on all his beliefs (thanks to Definition 1). It yields (default) constraints on G from intentions: if one knows $\mathcal{I}_a\psi$ and nothing about G or about a's beliefs, then the minimal CP-net G that satisfies the default consequence is simply the global preference $\psi \succ_a \neg\psi$. As agents converse, each dialogue action may reveal new information about intentions, and via Maximising Utility part (b) this imposes new constraints on G. But while Maximise Utility part (b) is conservative about exactly which of a's beliefs his preference for ψ is conditioned on, his dialogue moves can reveal more precise information—e.g., the utterance *I want to go to the mall to eat* should be sufficient to infer *eat* : *mall* $\succ_i$ $\neg$*mall*. A detailed algorithm for extracting preferences and dependencies among them from conversation is detailed in Cadilhac et al. (2011), but the details of this aren't relevant for our purposes here.

Basic cooperativity follows from an axiom that makes all agents intend that their commitments be shared among all the other dialogue agents:

Intent to Share Commitment:
$(b \in D \wedge \mathcal{P}_{a,D}\phi \wedge \neg\mathcal{P}_{b,D}\phi) > \mathcal{P}_{a,D}\mathcal{I}_a\mathcal{P}_{b,D}\phi$

If a commits, when addressing b (among others), to content ϕ and b hasn't committed to this yet, then normally a is also committed to intending that b so commit. This rule captures basic cooperativity because b committing to a's commitments entails he *understands* a's commitments (Clark, 1996). Indeed, it captures something much stronger than basic cooperativity—an intention that your contribution be *accepted* by others. While this is stronger than basic cooperativity, we think it's *rational* even in non-cooperative dialogue contexts: why commit to content if you don't intend that others accept the commitment? In addition, we regiment a constraint on assertions proposed by (Perrault, 1990, p180), by refining this axiom for assertions: when a's address to b commits him to an assertion $\mathcal{K}$, then normally $\mathcal{P}_{a,D}\mathcal{I}_a\mathcal{B}_b\mathcal{K}$.

Now let's examine more carefully the special case of Gricean cooperativity. We start by defining a Grice Cooperative game:

Definition 2 *A game is* **Grice Cooperative (GC)** *just in case for any of its players a and b*

1. their speech acts normally have their con-

ventional purpose (e.g., they normally ask a question so as to know a true answer); and

2. $(\phi : \psi \succ_a \neg\psi) > (\phi : \psi \succ_b \neg\psi)$
(i.e., the agents' preferences normally align).

We can now prove all the axioms in Fact 1.

Fact 1 Sincerity: $(\mathcal{P}_{a,D}\phi \wedge GC) > \mathcal{B}_a\phi$
Sincerity for Intentions:
$(\mathcal{P}_{a,D}\mathcal{I}_a\phi \wedge GC) > \mathcal{I}_a\phi$
Sincerity for Preferences:
$(\mathcal{P}_{a,D}(\phi : \psi \succ_a \neg\psi) \wedge GC) > \phi : \psi \succ_a \neg\psi$
Competence:
$(\mathcal{P}_{a,D}\phi \wedge \mathcal{P}_{b,D}?\phi \wedge a, b \in D) \rightarrow$
$((\mathcal{B}_b\mathcal{B}_a\phi \wedge GC) > \mathcal{B}_b\phi)$
Cooperativity:
$(b \in D \wedge \mathcal{P}_{a,D}\mathcal{I}_a\phi \wedge GC) > \mathcal{I}_b\phi$

These axioms make any *declared* belief, intention or preference in a GC conversation normally an *actual* belief, intention or preference too (cf. the Gricean maxim of Quality (Grice, 1975, p45)). Competence makes belief transfer the norm (if b asked whether ϕ). This default likewise follows from Grice's Maxim of Quality as he described it in (Grice, 1989, p371): he stipulates that in order to contribute to a conversation via the Maxim of Quality, one must say what is true. To do otherwise is not to contribute inferior information; rather, it contributes no information at all. Furthermore, Lewis (1969) argues persuasively that unless such a principle of competence forms the basis of cognitive modelling, then one cannot construct a sound philosophical argument that explains why linguistic conventions come into being in the first place, or why we assume that a speaker whom we understand is speaking the same language as we are—a hallmark of basic cooperativity. Finally, Cooperativity makes a declared individual intention normally a shared actual intention (recall the Gricean notion of utterance meaning in conversation (Grice, 1969, p151) and the corresponding notion of Strong Cooperativity from Section 1). Such principles of sincerity and cooperativity are usually taken as primitive axioms in BDI approaches to dialogue; here, we *derive* them when agents D are players in a joint game G that satisfies Definition 2.

Outline Proofs: Sincerity: Suppose $\mathcal{P}_{a,D}\phi$ and GC hold and moreover that ϕ expresses a proposition that is capable of being believed. Then we'll show that if all the normal GC consequences hold (see Definition 2), then $\mathcal{B}_a\phi$ must also hold.

By Intent to Share Commitment, $\mathcal{P}_{a,D}\mathcal{I}_a\mathcal{B}_b\phi$ defeasibly follows from our premises for any $b \in D$. By Maximising Utility and the fact that $\mathcal{I}$ is a D modality, $\mathcal{I}_a\mathcal{B}_b\phi$ defeasibly implies $\mathcal{B}_b\phi \succ_a \neg\mathcal{B}_b\phi$. Upon learning of a's commitment and the fact that the game is GC (in particular, clause 1 of Definition 2 means that the preference underlying a's move ϕ that we have just derived is a's actual preference), we infer $\mathcal{B}_b\phi \succ_b \neg\mathcal{B}_b\phi$. Assume further that belief preferences pattern after factual preferences. That is:

$$(\mathcal{B}_b\phi \succ_b \neg\mathcal{B}\phi) \rightarrow (\phi \succ_b \neg\phi)$$
$$(\neg\mathcal{B}_b\phi \succ_b \mathcal{B}\phi) \rightarrow \neg(\phi \succ_b \neg\phi)$$

So $\phi \succ_b \neg\phi$. Now suppose that $\neg\mathcal{B}_a\phi$. Then assuming we prefer our belief actions when we have them, $\neg\mathcal{B}_a\phi \succ_a \mathcal{B}_a\phi$, and therefore $\neg(\phi \succ_a \neg\phi)$. Thus the game cannot be a normal GC game, contrary to our assumptions. So $\mathcal{B}_a\phi$. Now, **Weak Deduction** is a valid rule of the weak conditional $>$ (Asher, 1995): if $\Gamma, \phi \mathrel{|\!\sim} \psi$, $\Gamma \not\mathrel{|\!\sim} \psi$ and $\Gamma \not\mathrel{|\!\sim} \neg(\phi > \psi)$ then $\Gamma \mathrel{|\!\sim} (\phi > \psi)$. So Weak Deduction yields the desired $>$ statement, $(\mathcal{P}_{a,D}\phi \wedge GC) > \mathcal{B}_a\phi$. □. We can also derive (though we don't show it here) a stronger version of Sincerity where a doesn't believe alternatives to what he said, yielding scalar implicatures.

Sincerity for Intentions: Suppose $\mathcal{P}_{a,D}\mathcal{I}_a\phi \wedge GC$. By Sincerity (which we've just proved), $\mathcal{B}_a\mathcal{I}_a\phi$. Since intentions are doxastically transparent (i.e. $\mathcal{B}_a\mathcal{I}_a\phi \leftrightarrow \mathcal{I}_a\phi$), the result follows with an application of Weak Deduction. □.

Sincerity for Preferences is proved in a similar way, using also the assumption that preferences are doxastically transparent. □.

Competence: Suppose $\mathcal{P}_{b,D}?\phi \wedge \mathcal{P}_{a,D}\phi \wedge b \in D \wedge \mathcal{B}_b\mathcal{B}_a\phi$ and a GC game. Given Definition 2, the intention that normally underlies asking a question (i.e., to know an answer) and Maximising Utility ensures that b's asking ϕ implies $\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi \succ_b \neg(\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi)$. So by GC

(i.e., the agents' preferences normally align), we also have: $\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi \succ_a \neg(\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi)$. By Maximising Utility we can assume that b's asking a question together with a's response are both optimal moves in equilibrium. These moves then should realise the preference $\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi \succ_b \neg(\mathcal{B}_b\phi \vee \mathcal{B}_b\neg\phi)$. Furthermore, by Sincerity, $\mathcal{B}_a\phi$. There are two choices now: either a is trustworthy or not. If a is not trustworthy, then his commitment to ϕ is no indication of its truth. But then there is a move (do not ask a whether ϕ) that would have been more advantageous for b (given that listening to someone and processing the response is a cost). So given that $\mathcal{P}_{b,D}?\phi$ is the equilibrium move—in other words, this is a move that is optimal for b in that it maximises his expected utility—b must believe a to be trustworthy, and so $\mathcal{B}_b\phi$. Using Weak Deduction thus yields Competence. □.

Cooperativity: Assume $b \in D \wedge \mathcal{P}_{a,D}\mathcal{I}_a\phi \wedge GC$. By Sincerity for Intentions, we have $\mathcal{I}_a\phi$. By Maximising Utility, we can infer *CP-solution*$_a(\phi, G)$, where G is the GC game with at least a and b as players. By GC and Competence, this defeasibly entails *CP-solution*$_b(\phi, G)$. And so Maximising Utility yields $\mathcal{I}_b\phi$. Using Weak Deduction gets us the desired $>$ statement. □.

Intention and belief transfer in a GC conversation is a *default*: even if preferences align, conflicting beliefs may mean agents have different CP-solutions making their intentions different too (by Maximising Utility), and Competence may apply but its consequent isn't inferred. Thus rejection and denial occur in GC dialogues (see (3)). On the other hand, in GC environments interpretations are normally credible: e.g., by Sincerity and Competence B's assertion (3b) yields belief transfer that there's no parking. This is a simple, symbolic counterpart to the much more elaborate result concerning credibility from Crawford and Sobel (1982).

4 Related Work

In contrast to Gricean formalisations in BDI logics, we have conditioned Gricean behaviour on shared *preferences* rather than shared *intentions* (see Definition 2) and we have derived Gricean axioms from a more general axiomatisation of human behaviour rather than treating them as primitive.

Signalling games provide a basis for predicting conversational implicatures (e.g., Parikh (2001), van Rooij (2004)) and also insincerity—the less aligned the preferences, the less credible the signals (Crawford and Sobel, 1982). But signalling models either take a signal to mean whatever it is optimal for it to mean (thereby bypassing linguistic convention) or the mapping $[\![.]\!]$ from signals to meaning is fixed and monotonic (e.g., Farrell (1993), Franke (2010)), with pragmatic interpretations being entirely epistemic in nature: they arise when the optimal interpretation of s is distinct from $[\![s]\!]$. Our model differs in its view of *conventional meaning*: while we acknowledge that some pragmatic inferences are epistemic (e.g., see **Sincerity**), we also believe that $[\![s]\!]$ goes beyond lexical and compositional semantics because it is constrained to be *coherent* (Lascarides and Asher, 2009). But this makes computing $[\![s]\!]$ defeasible, which reflects the fact that all inferences about coherence are defeasible. So in non-cooperative conversation, an interlocutor must test rigorously his defeasible inference about what the speaker is publicly committed to, as well as test the credibility of that commitment (i.e., whether the speaker believes it). We hope that CL can model such tests, but leave this to future work.

5 Conclusions

We have proposed a qualitative model of cognitive reasoning with several desirable features for modelling dialogue: it supports reasoning with partial information about preferences; and it distinguishes the public commitments one makes through utterances and private mental states that affect and are affected by them. The axioms of the cognitive logic approximate rational action from game theory and compel agents to be basic cooperative. We showed that Gricean principles of sincerity and cooperativity are derivable from them when the agents' preferences nor-

mally align.

We have focused here entirely on the cognitive model; linking it to dialogue content is ongoing work. The cognitive logic should also be dynamic since dialogue actions trigger changes to mental states: our static CL can be made dynamic with no cost to complexity by exploiting public announcement logic (Asher and Lascarides, 2011). Finally, progress in analysing strategic conversation requires an extensive study of data in many domains: e.g., political debate, commercial negotiations, courtroom cross examination and others. The *Settlers* dialogues cited here are all taken from our ongoing corpus collection effort, in which utterances are aligned with machine readable game states. We hope to release this corpus, labelled with rich semantic and cognitive information, in due course.

Acknowledgements: This work is supported by ERC grant 269427 (STAC).

References

CE Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meeting contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

N. Asher. Commonsense entailment. In G. Crocco, L. Farinas, and A. Herzig, editors, *Conditionals: From Philosophy to Computer Science*, pages 103–145. OUP, 1995.

N. Asher. Implicatures in discourse. to appear in *Lingua*, in press.

N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

N. Asher and A. Lascarides. Commitments, beliefs and intentions in dialogue. In *Proceedings of Londial*, pages 35–42, 2008.

N. Asher and A. Lascarides. Reasoning dynamically about what one says. *Synthese*, 183(1):5–31, 2011.

E. Bonzon. *Modélisation des Interactions entre Agents Rationnels: les Jeux Booléens*. PhD thesis, Université Paul Sabatier, Toulouse, 2007.

C. Boutilier, R.I. Brafman, C. Domshlak, H.H. Hoos, and David Poole. Cp-nets: A tool for representing and reasoning with conditional *ceteris paribus* preference statements. *Journal of Artificial Intelligence Research*, 21:135–191, 2004.

A. Cadilhac, N. Asher, F. Benamara, and A. Lascarides. Commitments to preferences in dialogue. In *Proceedings of SIGDIAL*, pages 204–215, 2011.

H. Clark. *Using Language*. Cambridge University Press, Cambridge, England, 1996.

V. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

J. Farrell. Meaning and credibility in cheap-talk games. *Games and Economic Behaviour*, 5:514–531, 1993.

M. Franke. Semantic meaning and pragmatic inference in non-cooperative conversation. In T. Icard and R. Muskens, editors, *Interfaces: Explorations in Logic, Language and Computation*, pages 13–24. Springer-Verlag, 2010.

H.P. Grice. Utterer's meaning and intentions. *Philosophical Review*, 68(2):147–177, 1969.

H.P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 41–58. Academic Press, 1975.

H.P. Grice. *Studies in the Way of Words*. Harvard University Press, Cambridge, Massachusetts, 1989.

B. Grosz and C. Sidner. Plans for discourse. In J. Morgan P. R. Cohen and M. Pollack, editors, *Intentions in Communication*, pages 417–444. MIT Press, 1990.

C. Hamblin. *Imperatives*. Blackwells, 1987.

A. Lascarides and N. Asher. Agreement, disputes and commitment in dialogue. *Journal of Semantics*, 26 (2):109–158, 2009.

D. Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969.

P. Parikh. *The Use of Language*. CSLI Publications, Stanford, California, 2001.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffmann, 1988.

R. Perrault. An application of default logic to speech act theory. In J. Morgan P. R. Cohen and M. Pollack, editors, *Intentions in Communication*, pages 161–186. MIT Press, 1990.

K. Schulz. *Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals*. PhD thesis, University of Amsterdam, 2007.

R. van Rooij. Signalling games select horn strategies. *Linguistics and Philosophy*, 27:493–527, 2004.