Strategic Conversation

Nicholas Asher, Alex Lascarides

Strategic conversation*

Nicholas Asher
CNRS, Institut de Recherche en Informatique de Toulouse

Alex Lascarides
School of Informatics University of Edinburgh

Abstract Models of conversation that rely on a strong notion of cooperation don't apply to *strategic conversation* — that is, to conversation where the agents' motives don't align, such as courtroom cross examination and political debate. We provide a game-theoretic framework that provides an analysis of both cooperative and strategic conversation. Our analysis features a new notion of *safety* that applies to implicatures: an implicature is safe when it can be reliably treated as a matter of public record. We explore the safety of implicatures within cooperative and non cooperative settings. We then provide a symbolic model enabling us (i) to prove a correspondence result between a characterisation of conversation in terms of an alignment of players' preferences and one where Gricean principles of cooperative conversation like Sincerity hold, and (ii) to show when an implicature is safe and when it is not.

Keywords: non-cooperative conversation, implicature, discourse coherence, game theory, cognitive modelling

1 Introduction

A theory of dialogue should link discourse interpretation and production to general principles of rational action and decision-making. Grice (1975) and his followers provide a theory meeting this constraint when conversational actions obey a strong cooperative principle. On this view, agents

* Thanks to participants at SIGDIAL 2008 and 2011, SEMDIAL 2008, and SuB 2008, to David Beaver, Ron Petrick, Chris Potts, Mark Steedman, Matthew Stone, Rich Thomason, Kai von Fintel and anonymous reviewers for comments. Any mistakes that remain are entirely our responsibility. This work is supported by ERC grant 269427.

cooperate on a level of intentions: once an agent in a conversation learns of someone's conversational goals he either adopts them and attempts to realise them, or he says why he can't adopt them. In technical work, this principle is sometimes expressed in terms of aligned utilities (van Rooij 2004, Lewis 1969). This principle leads agents to coordinate on the conventions that govern linguistic meaning (basic cooperativity) and on intentions about conversational goals, what we'll call *Gricean cooperativity*.

But Gricean cooperativity is a hallmark of only some kinds of conversation. Consider the cross-examination in (1) of a defendant by a prosecutor, from Solan & Tiersma 2005 (thanks to Chris Potts for this example):

(1) a. P(rosecutor): Do you have any bank accounts in Swiss banks, Mr. Bronston?

b. B(ronston): No, sir.

c. P: Have you ever?

d. B: The company had an account there for about six months, in Zurich.

The locutionary content of (1d) is true. But Bronston deflects the prosecutor's enquiry by exploiting a misleading implicature, or *misdirection*: (1d) implicates that Bronston never had a Swiss bank account and this is false. As Solan & Tiersma (2005) point out, what's a matter of public record, what a speaker publicly commits to, can be a matter of debate. We agree with their judgement that Bronston's implicature is plausibly part of the public record but that he can also defensibly deny that he publicly committed to that answer.

Misdirections happen outside the courtroom too. Consider dialogue (2) uttered in a context where Janet is trying to get Justin, who is the jealous type, off her back about her old boyfriend Valentino (from Chris Potts (pc)).

(2) a. Justin: Have you been seeing Valentino this past week?

b. Janet: Valentino has mononucleosis.

Suppose that Valentino indeed has mononucleosis. Janet's response implicates that she hasn't seen Valentino, an implicature which we think holds even if she has seen him and/or Justin isn't confident that Janet shares his intentions to get a true and informative answer to his question (2a). But as with (1d), Janet can defensibly deny that she committed to the negative answer.

Dialogues like (1) and (2) exhibit what we call *rhetorical cooperativity*, but

not Gricean cooperativity. A rhetorically cooperative move is a speech act one would expect from a speaker who fully cooperates with his interlocutor. Rhetorical cooperativity makes a speaker *appear* to be Gricean cooperative although he may not actually be so. This is a frequent feature of strategic conversations, in which agents' interests do not align. Here are some examples of rhetorically cooperative speech acts: a rhetorically cooperative response to a question implies either an answer or that the respondent doesn't know the answer; a rhetorically cooperative response to an assertion either implies that the assertion is grounded or it identifies why it's not (Clark 1996); a rhetorically cooperative response to a greeting returns the greeting. It might seem that responses follow from and in turn suggest Gricean cooperativity; nevertheless, rhetorical cooperativity and Gricean cooperativity are, as we will argue, independent notions.

In this paper, we explore rhetorical cooperativity, which we take to be a basic phenomenon, in the absence of Gricean cooperativity. In Section 2 we formally characterise Gricean cooperativity and sketch how these principles can yield implicatures. We then show that our Gricean theory has too few tools to model the implicatures that we and others (Solan & Tiersma 2005) claim are drawn from (1d) or (2b). We formulate in Section 3 a general theory of conversation that's based on *preferences* and *discourse coherence*, which we implement in Section 4 in a classic game theoretic framework. This models rhetorical cooperativity in detail, accounts for the implicatures in (1d) and (2b), and highlights an important difference between Gricean cooperative conversations and non cooperative ones. The difference between a Gricean cooperative conversation and a non cooperative version of the same exchange lies not in the implicatures that can be drawn from it, as the Gricean would have it, but in what we'll call the *safety* of the implicatures: the reliability of inferences about what message a speaker conveyed. Despite the rigour and clarity of the game-theoretic model, it tells us what agents should do but doesn't tell us anything about the mistakes they make or how they actually reason about conversation. To rectify this, we provide in Sections 5 and 6 a proof-theoretic counterpart to the game theoretic model where we sketch how resource-bounded agents might reason about conversation. We provide principles that allow us to infer (i) when inferences about conversation are safe, and (ii) that Gricean cooperativity is equivalent to an alignment of the agents' preferences (proofs are in the Appendix).

2 Problems for Gricean analyses of strategic conversation

In cooperative conversation, people normally believe what they say and help other agents achieve the goals they reveal through their utterances. Following Asher & Lascarides (2003), we offer below one way to make some of Grice's (1975) maxims a bit more precise, by appealing to defeasible generalisations.

- Rationality: normally, people act so as to realise their intentions.
- Sincerity: Normally agents who say ϕ should believe ϕ.
- Quantity: Normally, agents say as much as they need to, given their conversational goals.
- Competence: Normally, if B believes that A believes ϕ, then B should believe ϕ.
- Sincerity about Intentions: Normally if A M-intends ϕ (that is, A's utterance implies that A intends ϕ and intends that this intention be recognised), then A intends that ϕ.
- Strong Cooperativity : Normally, if A M-intends that ϕ, then B should intend that ϕ.

It is important to understand what these defeasible generalisations say and don't say. Strong Cooperativity, for example, doesn't say that the hearer's recognition of an M-intention or conversational goal *always* leads the hearer to adopt this intention; it says that in the absence of factors to the contrary, the hearer will do so.

We also assume that additional defeasible rules link speech acts to intentions: for instance, if an agent asks a question, then normally he M-intends to know the answer to it (that is, this speech act implies that he intends to know an answer, and intends that this intention be recognised). The above default rules are proposed purely for the sake of concreteness; for a different but largely equivalent approach see e.g., Schulz 2007.

Such rules provide the basis for inferring implicatures, including *inter alia* scalar implicatures. These inferences are of the form that in saying u, perhaps in contrast to some set of alternative utterances that the speaker could have said but didn't, one infers that the speaker meant p, where p goes beyond the compositional and lexical semantics of u. Consider (3):

(3) a. A: Did all of the students pass?

b. B: Some passed.

Most interlocutors would infer from (3b) that *B* does not believe that all the students passed. Here's how this inference would follow given our precisification of Gricean principles. Suppose that the chosen move, in this case (3b), conforms to all the Gricean principles given above. Suppose further that alternative moves that the speaker could have chosen but didn't (normally) deviate from at least one of those constraints. Suppose also that it is given, either by discourse structure or by the lexicon, that the set of alternatives to *some* in (3b) is {none, some, all}. Then an informal sketch of the derivation of the scalar implicature proceeds as follows:

- Sincerity: (defeasibly) implies *B* believes his response to *A*'s question (and so believes that the alternative move using *none* would deviate from Sincerity). Competence (defeasibly) implies that *A* should believe this response; i.e., that some students passed.
- Strong Cooperativity: *B* intends *A* to know an answer to his question — that either all the students passed or they didn't.
- Rationality: *B*'s response realises that intention; i.e., it provides *A* with an answer.
- B *didn't* say *all the students passed*, which would have implied an answer. So choosing this alternative would not have deviated from Strong Cooperativity or Quantity; it must therefore deviate from Sincerity. So *B* doesn't believe that all the students passed.

If in addition *A* believes that *B* knows an answer (as is often the case when *A* asks a question as part of a plan to find out an answer), then the implicature that *B* doesn't believe that all students passed yields a stronger inference, that *B* believes not all of them did, which by Competence *A* will (defeasibly) adopt as his own belief too.

Rather than fully formalise this reasoning in a particular nonmonotonic logic (see Asher 2013 or Schulz 2007 for some proposals), let's step back and consider its essential features. Specifically, it requires Strong Cooperativity to derive scalar implicatures — that interlocutors normally adopt each other's conversational goals and that speakers tell the truth. Strong Cooperativity makes precise what we mean by cooperativity at the level of *intentions*: once

an agent in a conversation learns of someone's conversational goals he either adopts them and attempts to realise them, or he says why he can't.

But conversations can have purposes that deviate from strong cooperativity: people talk to bargain, to bluff, to mislead, to show off or promote themselves, to put others down. Such purposes lead people to misdirect and to conceal information that's crucial to the achievement of conversational goals. Moreover such moves crucially exploit the fact that participants in such conversations draw implicatures. But without Strong Cooperativity, the above Gricean derivation fails: one can't draw the scalar implicature.

This is directly relevant to our examples (1) and (2). Most competent speakers, including the prosecutor himself, interpret Bronston's response as appealing to a scalar implicature that provides a (negative) answer to the question.[1] Indeed, the prosecutor and others interpret Bronston's move this way, *even if* they believe the implicature is false — that is, they believe or have evidence that Bronston did have an account. The prosecutor wouldn't have tried (and convicted) Bronston for perjury if he didn't have evidence that contradicted the negative answer that he took Bronston as providing in the trial.

But modelling this scalar implicature via just our Gricean principles doesn't work in this context: if an interpreter believes that Bronston had an account (as the prosecutor did), then the consequent to the defeasible rule Competence cannot be inferred. Furthermore, the prosecutor would infer from his own belief that Bronston had an account that Bronston believes he had an account (on the grounds that people have complete and accurate information about the bank accounts they have); in other words, the prosecutor would infer that Sincerity doesn't fire either. But then rationality would dictate that Bronston wouldn't have intended (1d) as a true and informative answer to (1c), and this would follow only if he didn't adopt the prosecutor's intention to provide the prosecutor with a true and informative answer. So the situation is one where the consequent of Strong Cooperativity cannot be inferred, and therefore there is no way to derive the implicature in this situation.

If we rely only on our Gricean principles to compute implicatures, we predict that *in this context* Bronston's response is no different from one that explicitly evades the question (e.g., *I refuse to answer*) or asserts some random fact (e.g., *I'm six feet tall*). Since Strong Cooperativity doesn't hold, there is no way to run the defeasible reasoning that turns the response into

1 Solan & Tiersma (2005) argue for this position.

an indirect answer. That seems to us to be the wrong prediction. Anyone who is a competent speaker construes Bronston's response (1d) as being much closer to an answer than an utterance of a random fact about Bronston, like *I'm six feet tall*. It's also not an explicit evasion of the question, like *I prefer not to answer the question*.

The same problem applies to a Gricean analysis of (2). Our assumptions about Justin's suspicions entail that he must assume that he and Janet are in a non-cooperative or strategic situation. Nevertheless intuitively, even though Justin assumes that Janet won't adopt his intention to know an answer, Justin, and we, take Janet's response to be an indirect answer to his question. To obtain this interpretation, Justin must engage in some defeasible reasoning to connect the response (2b) with the question (2a). Note that this doesn't entail that Justin accept the answer or find it credible: we are interested only in what information Justin extracts from it. Of course, Janet can argue she is committed only to the factual content of her claim. But we assume that her choice to utter (2b) is based on the fact that as a competent conversationalist, she realises that (2b) is naturally interpreted as a (negative) indirect answer. She picks that response as one that best realises her goals, because she knows that interpreting (2b) as an answer to (2a) involves some defensible, but defeasible, reasoning on Justin's part, and that she has the option of denying that his reasoning is sound in the present case or that she was completely responsible for it. In any case, for both Justin and Janet, the consequence of Strong Cooperativity isn't inferable. Thus a Gricean can't generate the implicature as part of the interpretation of (2b). So contrary to intuitions, a Gricean must predict that Janet's response, like Bronston's, is no different from an assertion of some random fact.

Notice that misdirection is quite different from lying. Suppose that Janet had lied by saying *No* to Justin's question (2a). The linguistic meaning of the answer, assuming just rhetorical cooperativity, would be clear: *No* means that Janet hasn't been seeing Valentino. Similarly for Bronston's direct answer in (1b). Intuitively this case is different from the misdirections in (1d) and (2b). The message in this discourse context is unambiguous and direct.

Our argument crucially relies on Strong Cooperativity, and we suspect that many Griceans, e.g., Green (1995), will think it is too strongly worded. Green argues that one should adopt no more of the intentions of the questioner than is required for the conversational purpose, as Grice himself said. He adds as a separate stipulation that the asking of a question invokes the conversational purpose that the recipient give "a complete answer to the question if they

can" (Green 1995: p.107). But whose conversational purpose are we talking about? Presumably the questioner's. Without Strong Cooperativity, it doesn't follow that the recipient will adopt this conversational purpose. Without the adoption of this purpose, we can't infer that the recipient is answering the question. Green and many other Griceans overlook the centrality of Strong Cooperativity in a Gricean derivation of implicatures.

A Gricean might suggest that the difference between an assertion of just any random fact and Janet's actual response in (2) is the *counterfactual claim* that had they been in a cooperative situation, Janet's response would have been an indirect answer, while an assertion of some random fact would not. Nevertheless, it's hard to see what this counterfactual claim does for the interpretation of Janet's response *in the actual context*. In the counterfactual Gricean cooperative context, Justin, we agree, would draw the implicature, using the sort of reasoning we outlined above and the defeasible conclusions of Strong Cooperativity. In the actual context, however, Justin is the jealous type and so suspects that Janet doesn't intend to provide a truthful answer. He doesn't believe that the consequent of Strong Cooperativity holds. Nevertheless, he *still* draws the implicature; that's why Justin will be justified in being angry *in the actual context*, if he finds out Janet has been seeing Valentino. Justin's jealousy and his suspicions make him wary of the indirect answer. However, not believing or being wary of the indirect answer is a matter of *credibility* and *belief* about the information imparted; it's *not* a matter of what information is imparted by the response. Griceans have no way of explaining why this is the case. For them it is irrational to draw the implicature in the actual context, and it would be irrational for Justin to be angry.[2]

An alternative constraint besides Strong Cooperativity would be to assume that speakers make their contribution to the dialogue *relevant* to the dialogue so far (cf. Grice's maxim of Relevance). But this maxim is too weak to generate the implicature on its own. To see why, consider replacing the inference step where Strong Cooperativity is applied in our derivation of the scalar implicature for (3b) with a maxim of relevance. Then it does *not* follow that *B*

2 A Gricean might argue that implicatures in the courtroom case arise from an exogenous constraint on conversation like the oath to tell the truth, the whole truth and nothing but the truth in a courtroom. To derive scalar implicatures from the oath, however, the Gricean must stipulate that "nothing but the truth" entails all scalar implicatures. But this is far too rigid an interpretation: if the oath did force all implicatures to be entailed, then the Supreme Court would have had no business overturning Bronston's conviction for perjury, which in this case it did — in effect it would turn all implicatures into entailments.

feels compelled to supply an answer: a partial answer, or some other relevant contribution such as a commentary on the question, would also be relevant. Similarly, in (1d), a relevance maxim may distinguish this utterance from an assertion of a random fact (in that (1d) is about holding Swiss bank accounts, while being six foot tall has nothing to do with bank accounts). But it would *not*, on its own, be enough to infer that (1d) implicates an *answer*: other relevant semantic relations such as being a commentary on the question would comply with the defeasible relevance constraint too.

We suspect that Gricean principles are not sufficient in strategic contexts even to identify *direct* answers containing anaphoric elements as answers. To interpret (1b) as a direct answer, we first have to assume it is somehow related to (1a). Theories of discourse structure, like Question Under Discussion (QUD, Ginzburg (2012)) or Segmented Discourse Representation Theory (SDRT, Asher & Lascarides (2003)) study such attachment decisions and attempt to predict them. In such theories, an assumption that (1b) is attached to (1a) constrains the antecedents to anaphoric expressions — in (1), the anaphor is *no* and the assumption that it's attached to the question (1a) forces its antecedent to be the core proposition of (1a), thereby yielding its interpretation as a direct answer. The theory of attachment and constraints on the interpretation of anaphora in such discourse theories bypass entirely any reasoning about mental states or cognitive states in this example, and they are sufficient to infer an interpretation of (1b) as a direct answer to (1a).

Gricean inference doesn't exploit discourse structure in this way, relying instead on reasoning about *mental states* to augment the compositional semantics of an utterance to a specific interpretation. (1a) was not dialogue initial: there are many propositions in the context that could conceivably act as antecedents to the anaphor in (1b). So *even if* one assumes that (1b) is a response to (1a), then without the machinery of discourse structure being used to constrain the interpretation of anaphora, we don't see how a Gricean can identify the appropriate antecedent to the anaphor without appealing to Strong Cooperativity.

Strong Cooperativity can solve the problem we have just discussed. Consider (4) which allows us to infer that (1b) is a direct answer to (1a):

(4) If response X can be interpreted as an answer to question Y, then normally interlocutors interpret X as an answer to Y.

Strong Cooperativity entails (4): the respondent makes it his own goal that the questioner know a true answer to his question, and so by rationality the respondent's action is construed to be a part of a plan to satisfy that goal, making an interpretation of *no* as a direct answer predictable, rational and natural. But without Strong Cooperativity, the derivation of (4) fails. Suppose that the prosecutor knew that Bronston did in fact have a bank account at the time of question — i.e., that he was explicitly lying. Then Competence and Sincerity fail, and as before the prosecutor concludes from this that Bronston doesn't share his conversational goals, which in turn undermines the derivation of the implicature in (4). So we can't even infer that (1b) is a direct answer to (1a), at least not using Gricean inference alone. In other words, using Gricean principles alone, one cannot infer that there is rhetorical cooperativity unless there is also Gricean cooperativity. This is completely counterintuitive: (1b) is clearly interpreted by everyone as a direct answer, even if Bronston is lying. This is not a matter of credibility — whether the response is believable — but in fact a problem of what information was communicated. Something needs to supplement the Gricean picture, or to replace it.[3]

A Gricean could retreat to the level of Bronston's intentions. A Gricean might claim that clearly Bronston has the intention of having the prosecutor believe that he has answered both questions and the prosecutor recognises that intention. But how is that intention recognised? It's not by cooperativity of intentions, for the reasons we have already given. Rather, the intention is inferred because of basic linguistic norms that function at a discourse level, regardless of cooperativity (Traum 1994, Poesio & Traum 1997, Traum et al. 2008). That is, the basic linguistic forms, together with the choice of which part of the discourse context the current contribution attaches to, affects content: it makes *B*'s response an answer in both (1b) and (1d). Implicatures thus flow from the discourse structure, rather than the other way round as in Grice.[4] These are linguistic norms that theories of discourse structure

3 Of course (4) doesn't entail Strong Cooperativity. We could use (4) on its own or other such principles to construct rhetorical cooperativity. We believe that Green (1995) has something like this view, and it also accords with our prior work (Asher & Lascarides 2003). But if one does this, the Gricean Principles become irrelevant for deriving rhetorical cooperativity, which we believe is as it should be.

4 For a detailed argument on this score, see Asher 2012, 2013.

try to capture — they facilitate meaning making by resolving the meaning of each communicative act through constrained inference over information that's made linguistically salient in the prior discourse. These norms are of course compatible with a Gricean view of communication. But they don't follow from a Gricean picture of communication as enshrined in the maxims, and they are also compatible with conversations which don't feature Gricean cooperativity.

To conclude, people often communicate in the context of diverging interests — any time you want to bargain for a car, invest in stocks, play a competitive game, get your parents to do something for you, your interests may fail to align with those of your conversational partners. Nevertheless, people continue to draw implicatures like those we've discussed. They draw these implicatures when cooperativity at the level of intentions is lacking, and the Gricean makes a false prediction if he takes implicatures to be generated entirely from cooperative principles associated with the maxims. So rather than force derivations of implicatures in strategic contexts into a model that is based on cooperativity of intentions, we will provide in the next sections an alternative foundation for deriving implicatures, which handles contexts where the agents' preferences and intentions diverge, as well as cases where they are aligned.

3 Our model

We hold that the following three features should be a part of any general model of conversation:

Public vs. Private: Speaking makes an agent *publicly commit* to some content (Hamblin 1987). A dialogue model must distinguish private attitudes from public commitments.

Coherence: Messages must be interpreted with respect to an underlying model of discourse coherence; likewise, decisions about what to say are influenced by coherence.

Levels of cooperativity: The model of dialogue must distinguish several levels of cooperativity, distinguishing at least rhetorical cooperativity from full Gricean cooperativity.

Insincerity motivates the **Public vs. Private** distinction. Many traditional mentalist models of dialogue based on Grice *equate* dialogue interpretation

with updating mental states. For instance, interpreting an assertion that p is equivalent to updating the model of the speaker's mental state to include a belief in p (e.g., Allen & Litman 1987, Grosz & Sidner 1990, Lochbaum 1998). But they are not equivalent in (1); we interpret Bronston's move (1d) as an indirect answer even if we know that Bronston believes the negation of its implied answer, making Bronston's beliefs inconsistent with the dialogue's interpretation. And we interpret (1b) as a direct answer to (1a) even if we know *B* believes he has a bank account.

The idea that a model of dialogue should distinguish a *public record* of discourse content from the private information about the agents' mental states isn't new: concepts such as common ground (Stalnaker 2002) have been proposed and motivated on the basis of several phenomena, including grounding (Clark 1996, Traum 1994), presupposition (Lewis 1969, Green 2000) and anaphora (Asher & Lascarides 2003). Insincerity provides another motivation. However, there is a link between the public record and private attitudes and to articulate this link, we say that speakers *publicly commit* to their contribution to the dialogue (Hamblin 1987, Traum 1994, Lascarides & Asher 2008). Thus, the principle of Sincerity says that a speaker normally believes what he publicly commits to.

Based on our discussion in Section 2, the fact that speakers commit to content that goes beyond that expressed by a sentence in isolation motivates the constraint of **Coherence**. For example, the fact that (1b) commits Bronston to an answer to (1a) entails a commitment to a coherent and specific interpretation of those utterances, with anaphoric expressions resolved to specific values. Models of discourse interpretation that exploit coherence assume that each discourse move bears at least one coherence relation to at least one other part of the discourse, and that resolving aspects of content like identifying antecedents to anaphora or resolving linguistic ambiguities depends on the discourse coherence (Hobbs 1979, Asher 1993, Kehler 2002). Coherence relations are an irreducible part of the content of discourse: the relations and the structure they engender on the discourse determine what's available for subsequent anaphoric reference. Asher & Lascarides (2003) thus include coherence relations as part of a discourse move in dialogue.

Coherence relations include *Explanation*, *Acceptance* (expressed by utterances like *I agree*) and *IQAP* (Indirect Question Answer Pair), the relation that holds between Bronston's response (1d) and the question in (1c) asked by the prosecutor. An indicative response to a question may be related to the question in several other ways, in a theory of discourse coherence like

SDRT (Asher & Lascarides 2003). It may be a direct answer, represented with the relation *QAP* (Question Answer Pair); or it may imply the respondent doesn't know the answer (e.g., *I don't know*), represented with the relation *NEI* (*Not Enough Information*). It may *deny* a presupposition of the question (the relation *Correction*) or it may reject the intention underlying the question (SDRT's *Plan-Correction*; e.g., *I refuse to answer*). It may elaborate a plan to know an answer while not implying a particular answer (SDRT's relation *Plan-Elab*), or it may provide a *Commentary* (e.g., *That's an interesting question.*). Different theories of discourse coherence posit different coherence relations, although all assume the set is finite (Hobbs 1979, Mann & Thompson 1987, Kehler 2002).

Each coherence relation has its own semantics. For instance, $QAP(a, b)$ entails that b is a direct answer to the question a according to the compositional semantics of questions and answers. $NEI(a, b)$ entails that b implies that its speaker doesn't know an answer to a. $IQAP(a, b)$ entails b defeasibly implies, via default rules that the questioner and respondent both believe, a direct answer to the question a. For instance, b may imply a direct answer via a quantity implicature. Indeed, the *only* way of making $IQAP(1c, 1d)$ consistent with the compositional semantics of (1c) and (1d) is to assume that (1d) implies a negative answer via a quantity implicature.[5]

To assume that a response is *always* coherently related to a prior turn is too strong. Had Bronston responded to (1c) with content that lacks any coherent connection (e.g., *I would like some water*), there would be no implicated answer and no perjury. Nevertheless, coherence is a key concept in defining public commitments in discourse: if there's a possible coherent link with previous discourse, interpreters will assume that this link is a part of the speaker's contribution to the dialogue's content — he is publicly committed to it and its consequences.

We need Coherence to analyse misdirections like that in (1). Bronston's utterance of (1d) commits him arguably to a negative answer to (1c); for us, this means that he's committed to a certain coherence relation holding between the question in (1c) and his own contribution in (1d), and to the semantic consequences of that relation.

5 Bronston cannot feasibly base his defence against perjury on a claim that he intended (1d) to imply a *positive* answer to (1c) — the semantics of *IQAP* demand that the answer be inferable using shared beliefs, and the proposition that employees normally have accounts at their company's bank is an unlikely individual belief let alone one that Bronston can argue is shared.

Asher & Lascarides (2003) provide axioms in what they call a *glue logic* (GL) to infer coherence relations that are part of the speaker's public commitments. The axioms in GL exploit information from lexical and compositional semantics as well as information about the discourse context, and they license defeasible inferences to a specific discourse interpretation, including rhetorical connections. An example GL axiom is (5), where $A > B$ means *If A then normally B*:

(5) $(\lambda :?(\alpha, \beta) \wedge qu(\alpha)) > \lambda : IQAP(\alpha, \beta)$

(5) states that if discourse segment β connects to segment α to form part of the content of the segment λ, but we don't know what that connection is (expressed by the GL formula $\lambda :?(\alpha, \beta)$), and α is a question (expressed by the GL formula $qu(\alpha)$), then normally β is an indirect answer.[6] Given an assumption that (1d) attaches to (1c), (5) defeasibly implies that they are linked by indirect answerhood, an interpretation that intuitively is salient and natural for (1d) (as the subsequent perjury conviction attests).

The presence of these two features in our model of conversation leads to two questions. First, what commitment has a speaker made with a particular contribution and is the commitment one that he can defensibly deny? And secondly, when do public commitments invite an inference to belief? Game-theoretic models of human behaviour show that preferences are of fundamental importance in answering the second question. Whether one should believe a speaker's commitments, what game theorists call *credibility*, is a function of the degree of overlap among the agents' preferences; the less the overlap, the less secure the inference (Crawford & Sobel 1982). We'll show in Section 5.2 that alignment of preferences is equivalent to our formalisation of Gricean cooperativity. But we don't explore credibility further here. Instead, we limit our analysis to the first question. In non-cooperative situations, identifying *commitments* is also sometimes problematic and can be open to debate. While we assume fixed lexical meanings and compositional semantics, identifying particular coherence relations among utterances is a matter of defeasible inference (Hobbs et al. 1993, Asher & Lascarides 2003). Since speakers publicly commit to coherence relations linking their speech acts to other contributions, computing a speaker's public commitment is a product of defeasible inference.

6 Asher & Lascarides (2003) justified the rule by appealing to Gricean cooperative reasoning. Asher 2012 offers a justification, to which we'll appeal below, that holds in non-cooperative or strategic situations.

We are interested in the safety of the inferences about coherence and about commitments. When is an inference about what a speaker has publicly committed to a safe inference? When is it safe to assume that a particular piece of content is a part of the public record? When Gricean cooperativity is assumed, defeasible inferences about the rhetorical function of an utterance are safe — they may fail to be sound but only because the speaker or his interlocutor didn't pursue a rhetorically clear strategy in language production or interpretation. But in non-cooperative conversations there may be clear but unsafe inferences to the rhetorical function of an utterance. Discourse coherence makes an interpretation of (1d) as an indirect (negative) answer highly salient. Nevertheless, it is not safe to treat the (defeasibly implied) negative answer as a part of Bronston's commitments, for it was deniable. Bronston could subsequently say with some justification "I never *said* I didn't have a bank account".[7] The notion of safety thus introduces a new distinction concerning discourse content and implicatures. Safety carves out that part of discourse content that is undeniably part of the public record.

Inferences based on coherence are sometimes safe, even in non-cooperative situations. An alternative way of conveying an indirect answer in (1) would have sustained the perjury conviction, even though the indirect answer follows from principles of discourse coherence:

(1) d′. Bronston: Only my company had an account, for about six months, in Zurich.

(1d′) implies a negative answer that goes beyond compositional semantics. For one thing, it relies on an attachment decision — namely that it attaches as a response to the question in (1c). In this case, however, answering *yes* to a subsequent question *yes or no, did you ever have a Swiss bank account?* would intuitively contradict (1d′). Why is this? The meaning of *only* is context-sensitive: it presupposes a *set of alternatives* (e.g., Rooth 1992). Work on presupposition has established that binding presuppositions is generally preferred to simple accommodation (van der Sandt 1992) and that presuppositional binding is a sort of anaphoric binding (Geurts 1996). The relevant alternatives set for (1d′) is naturally interpreted as bound by a set

7 The Supreme Court in fact over-turned the initial conviction for perjury: it acknowledged that (1d) was misleading but ruled that the prosecutor should have pursued a follow-up question, forcing Bronston to commit to the answer *no*. In effect the Supreme Court ruling as well as Solan and Tiersma's (2005) discussion point to the conclusion we establish here theoretically: it is not always safe to treat implicatures in non-cooperative conversation as part of the public record.

of discourse available antecedents that includes Bronston; and the binding of the presupposition to a set including Bronston depends on the inference that (1c) and (1d′) are coherently connected (for discourse availability is determined by coherence connections in (1a–cd′) (Hobbs 1979, Kehler 2002, Asher & Lascarides 2003)). If this coherence connection and the binding are possible, then the speaker commits to it, thereby committing to (1d′) being an indirect answer to (1c). Notice, however, that the inference to a negative answer from (1d′) is *still defeasible*. The "natural interpretation" relies on an assumption of coherence and a choice of contextually given alternatives; these could be wrong, but not plausibly deniable.

We've shown that all that's needed for interpreting (1d′) as an indirect (negative) answer, besides compositional and lexical semantics, is an assumption that it is coherent. This assumption is needed to resolve the anaphor *only*; it forces the value of *only* to one that makes (1d′) imply that Bronston didn't have an account. In contrast, interpreting (1d) as an indirect answer requires something more, *in addition* to the assumption that its interpretation is coherent. It does *not* follow from assuming that (1d) is coherent that it implies a negative answer. The compositional and lexical semantic content associated with (1d) is consistent, for instance, with other coherence relations. Given (1d)'s fixed meaning — i.e., its compositional semantics plus the assumption that Bronston's commitments contain coherence relations — it's possible, though at first glance not likely, that Bronston's commitment is to *Continuation*(1*b*, 1*d*), *Continuation* being a relation that entails its arguments address a common, contingent topic (e.g., Swiss bank accounts). This coherence connection does *not* commit Bronston to an answer to (1c). For reasons having to do with the maximisation of discourse coherence (Asher & Lascarides 2003), we choose the attachment point (1c) as the preferred attachment point, and then the defeasible principle (5) does its work, provided no information in the context blocks the derivation of the indirect answer. But the fact that intuitively, interpreting (1d) as an indirect answer isn't safe, while interpreting (1d′) is, suggests that coherence is a more basic and unavoidable commitment than those additional factors involved in interpreting (1d) as an indirect answer.

Notice that *either* interpretation of (1d) yields an implicature. The preferred interpretation of (1d) as an indirect answer gives one; its interpretation as a Continuation yields the conclusion that (1d) is not an answer to the question in (1c), though it is coherent. The puzzle is why is the latter implicature safe but not the former?

Finally, the contrast between (1d) and (1d′) separates the question of safety from that of credibility mentioned earlier: the negative answer implied by (1d′) is safe, but it may very well be incredible — i.e., the negative answer is a part of the public record, but shouldn't be believed. Again to underscore the difference between these concepts, credibility evaluates whether the speaker's public commitments can be believed. Safety, on the other hand, evaluates inferences that *identify* what the speaker's public commitments are. The identification of what is and what must be a matter of public record is logically prior to that of determining the credibility of those commitments.

We will analyse why inferences about commitments might break down when preferences aren't aligned. Following Asher 2012, we provide a game theoretic explanation that features politeness constraints of why interpreting (1d) as an indirect answer is natural and rational even in the absence of strong cooperativity. But it turns out that interpreting (1d) as committing Bronston to a negative answer is unsafe. Given the stakes, the prosecutor should have tested the interpretation for safety. This view differs from a Gricean one, which draws no implicatures in such a strategic situation at all. For the Gricean, coherence threatens to be abandoned as well: the upshot is that Bronston's message is simply noise. In Section 4 we propose a test for safety of the inference that a normal and salient interpretation (e.g., the interpretation of (1d) as an indirect answer) is a part of the public record in the light of other coherent interpretations (e.g., treating (1d) as an 'opting out' move or refusal to answer).

Our model distinguishes three levels of cooperativity: basic cooperativity at the level of compositional meaning, rhetorical cooperativity and Gricean cooperativity. This motivates the third feature of our model of conversation, namely **levels of cooperativity**. Dialogue (6) (from Chris Potts (pc)) is basic cooperative but not rhetorically or Gricean cooperative.

(6) a. R(eporter): On a different subject is there a reason that the Senator won't say whether or not someone else bought some suits for him?

b. S(heehan): Rachel, the Senator has reported every gift he has ever received.

c. R: That wasn't my question, Cullen.

d. S: The Senator has reported every gift he has ever received.

e. S: We are not going to respond to unnamed sources on a blog. (http://www.youtube.com/watch?v=VySnpLoaUrI)

(6) involves no misdirection; Sheehan simply refuses to engage with the question and neither implicates an answer nor that he doesn't know the answer.

4 Interpretation games

Game theory is a powerful tool for analysing strategic reasoning, in particular reasoning about communication. It can furnish a rationale for inferring when a respondent to a question would choose to provide an indirect answer, and it can also provide the rationale for undermining such an inference, even though the inference is consistent with the conventional semantics of the response and even a salient candidate interpretation (e.g., (1d)).

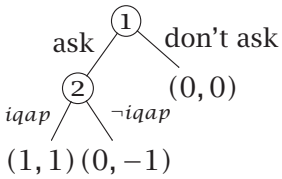
Game theory assumes that all rational agents act so as to maximise their expected preferences or utilities. We speak of *expected* preferences so as to handle the agent's uncertainty about the environment he's in; the agent may be uncertain about the outcome of an action, about other agents' mental states or what they might do, or any combination of these. Action choice in game theory is an optimal tradeoff between the likely outcome of an action you might choose to perform and the extent to which you want that outcome. With regard to dialogue, this view makes one's action choice (what to say for the speaker, how to interpret what's said for the interlocutor) dependent on one's expectations about the other agent's motives.

A game consists of a set of players, a set of actions available to each player and a mapping from each combination of players' actions to a real valued payoff or utility for each player. Typically, each player's payoff depends on *all* the players' actions, not just his own. A *Nash Equilibrium* (NE) is a combination of actions that is optimal in that no player would unilaterally deviate from it — in other words, each player expects at least as good a payoff from his own specified action than any other action that he could perform, assuming that all the other players do the actions that are specified for them in the NE.

Conversations are extensive and dynamic games: players typically perform their actions *sequentially* rather than simultaneously, with one player performing an action, and then another player performing an action. These sequential games are represented as trees: the leaf nodes capture each possible outcome of the game, and they are decorated with the utility of that outcome for each player (where a utility is a real number reflecting the extent to which the player prefers that outcome); each non-terminal node is labelled

with the player whose turn it is to act; and each arc is labelled with the action that the player whose turn it is performs. A sample sequential game tree is given in (7). Player 1 either asks a question or doesn't. If player 1 asks a question, then player 2 either responds with an indirect answer or responds with something that isn't an indirect answer; otherwise, player 2 says nothing.

(7)



The utilities for players 1 and 2 are given as pairs of integers. As is often the case in game theory, we assume that the game tree is common knowledge. In standard sequential games, players also know what actions the prior players performed during execution of the game.

Nash equilibria for simple games with perfect information, such as the one in (7), can be calculated using a procedure known as *backwards induction*: this identifies the subtree whose paths specify all and only those combination of actions that are NE. Roughly put, backwards induction proceeds as follows. Start at the leaf nodes. The preterminal nodes will be for player n, the last player in the game. If n is rational, she will choose an action that maximises her utility, given the state she's in (i.e., given what the prior players did). Graphically, this means that if a leaf node l_1 shares a mother with another leaf node l_2 with a higher utility for n, then delete that l_1 and its arc to the mother node. Now reason about the actions in this pruned tree of the player *n-1*, who plays before n: as before, *n-1* chooses the action that will ultimately maximise her utility after n plays, given what the previous players played. So one deletes any path from an *n-1* node to a leaf node with a lower utility than some other path that is rooted at that same *n-1* node. We continue to recursively prune branches all the way up the game tree, to yield a subtree of the original tree that's rooted at the first player's move and which consists of all and only those combination of actions that no one has an incentive to deviate from — the NE.

In our example (7), backwards induction starts with player 2's actions. Given the option of responding to a question, she will respond with *iqap* (because her utility for this is 1, compared with the alternative action ¬*iqap* with utility −1). So the arc labelled ¬*iqap* and its leaf with utility $(0, -1)$ is

pruned. If player 1 doesn't ask anything, then player 2 has no decision to make. Now player 1 has to decide what to do in the context of this pruned tree. Asking her question yields her a utility of 1 (for the $\neg$*iqap* branch is pruned); not asking yields her a utility of 0. So the "don't ask" branch is pruned. The resulting subtree thus captures just one NE: player 1 asks a question and player 2 responds with an indirect answer (*iqap*).

Recall that we are interested in the problem of safe *dialogue interpretation*: given the speaker's signal, which coherent interpretation of it, including resolved anaphora, is it safe to infer that the speaker is publicly committed to? So what we represented as a single action in the sequential game (7) becomes now a sequential game itself: a sequence of *two* actions where the speaker S performs some utterance u, and then the interlocutor or receiver R interprets it — i.e., he identifies the coherent logical form of u that S publicly committed to. As we showed in Section 3, public commitments aren't fully observable to R because they go beyond the compositional and lexical semantics of S's utterance. This combination of extensive form and uncertainty makes a dialogue interpretation game a special type of game that's known in the literature as a *signalling game*. In signalling games, the receiver does not know the sender's situation, and in particular what he intends to convey with a message. Solution concepts that are associated with signalling games provide the receiver with a rationale for choosing an optimal action in spite of this uncertainty. Likewise they provide a rationale for the speaker to choose his optimal signal (or what's known as his *message* in traditional signalling games), based on his estimates of how the receiver will respond to it.

More formally, a signalling game has a speaker S, a receiver R, a set of types T and a "dummy" player Nature who assigns a type to S (perhaps randomly drawn). Based on this type t, S chooses a signal $s \in S$ to send to R. R, on observing s, chooses an action $a \in A$ in response. While S knows his type, R doesn't. The game ends once R has issued his response action a. And both players have complete information about their own and the other player's relative preferences over the outcomes of the game. These preferences are represented as a utility function U_S and U_R from $T \times S \times A$ to the real numbers, making their preferences dependent not only on the actions they perform (i.e., S's signal s and R's reaction a), but also on S's type. A *strategy* for a player is a plan dictating the choices he will make in every situation he might find himself in. A (pure) strategy for the speaker S is thus a function μ from T to S; for R it is a function α from S to A.

Signalling games as they are traditionally construed assume that T defines the possible states of affairs. For example, in (3) there will be a state t_1 where no student passed, a state t_2 where some but not all students passed, and a state t_3 where all students passed. Given this conception of T, it is natural to model natural language interpretation using a signalling game where each of R's actions in A identifies a subset of T (e.g., van Rooij 2004, Franke 2010): in other words, $a \in A$ identifies R's take on the state of affairs that S intended to convey. For instance, if $\alpha(3b) = \{t_2\}$, then this interpretation includes the scalar implicature.

What we need as specified in Section 3 is a game where R identifies S's *public commitments*, with speakers committing to specific and coherent contents. So while the speaker strategies are functions μ from T to S as in standard signalling games, we take the signals $s \in S$ to have a particular linguistic form. The preferences U_S and U_R are, as before, made dependent on the speaker type and both players' actions: they are a function from $T \times S \times A$ to $\mathbb{R}$. Here we specify R's actions to involve the choice of a meaning representation that resolves s's linguistic ambiguities, anaphora and coherence relations to specific values. We call these *messages* and call the set of messages M. R's strategies are thus functions from signals in S to messages in M: on observing s, R chooses a message $m \in M$. The messages in M *and* the signals in S both have exogenous interpretations: $[\![.]\!]_{\mathcal{L}}$ is the semantic interpretation of signals given by the compositional and lexical semantics of the language (its value is a set of possible worlds); and $[\![.]\!]_{\mathcal{M}}$ is an assignment of possible worlds to elements of M. For any $s \in S$ and for any α of R, $[\![s]\!]_{\mathcal{L}} \supseteq [\![\alpha(s)]\!]_{\mathcal{M}}$. That is, the interpretation $[\![\alpha(s)]\!]_{\mathcal{M}}$ of the message $\alpha(s)$ is always at least as specific as the abstract and underspecified meaning $[\![s]\!]_{\mathcal{L}}$ that's assigned to s by the linguistic grammar. Or to put it another way, the range of actions available to R all abide by basic cooperativity, with the message $\alpha(s)$ that R infers being consistent with the compositional and lexical semantics of the signal s that S performed.

We also need to refine the conception of speaker types T for the application we have in mind. We distinguish the types in T from possible states of affairs, and *a fortiori* from the interpretation of messages or signals. S may choose to misdirect R about the message he sent, about what public commitments he made.[8] R thus needs to reason about whether S is likely to misdirect, or not. Thus, R hypothesises several different preferential profiles

8 This is similar to the bluffing in classic conceptions of bargaining games (Osborne & Rubinstein 1990).

for S — these are the $t \in T$ — over which he has a probability distribution. Each $t \in T$ conveys information about S and his *linguistic strategies* — e.g., whether or not he is prepared to misdirect when cross examined in court, whether or not he is the sort of person who is courteous and polite when he communicates, and so on. R's task, then, is to reason about the likely type of S, given what he observes (i.e., the signal s). His task in assessing the safety of his interpretation is to rigorously test his inference about what message m S declared by uttering s. A message m is *safe* if it satisfies such tests.

Let's look at our example (1) from this perspective, and discuss how to apply concept solutions for signalling games to inferences about safety. Figure 1 represents the prosecutor P's model of an interpretation game where Bronston (B) emits a signal s in response to P's question (1c), and P infers what message m B publicly committed to be uttering s. P assigns each outcome a utility for each player (shown in the order (P, B)). We assume that P thinks B is guilty — otherwise he wouldn't be prosecuting him. But P is uncertain about B's specific preferences in this game. Hence we postulate two player types each with their own utility function, where roughly speaking B_1 likes to misdirect P and B_2 doesn't (we'll shortly see that the utilities for B_1 and B_2 reflect this); so P's uncertainty about whether B wants to misdirect him is modelled via the probability distribution $Pr(\vec{B})$ over the random variable $\vec{B} = \{B_1, B_2\}$. As we said earlier, this probability distribution Pr over B's type reflects P's model of B's beliefs *and* it reflects a general model of human behaviour, capturing some generic assumptions about how speakers would behave in various circumstances. In particular, Pr is conditioned on general evidence about the extent to which humans like to misdirect their interlocutors in courtroom cross examination. We don't know what that theory of human behaviour might be, and so we simply make $Pr(\vec{B}) = \langle p, (1-p) \rangle$.

In Figure 1 P doesn't consider the possibility that B entertained any signal other than s_1 or s_2 — s_1 being the utterance (1d) that's actually performed and that apparently implicates a negative answer, and s_2 expressing a refusal to answer (e.g., *I refuse to answer*). The actions available to P, given the signals s_1 and s_2 and the semantics conveyed by their linguistic convention, are to interpret B's signal as an indirect answer (*iqap*) or as a refusal to answer, which we label here as $\neg$*iqap*. For instance, if the outcome of this game is $(s_1, iqap)$, then P has inferred that B has made a negative answer to (1c) a matter of public record (because B is publicly committed to that answer). If the outcome is $(s_1, \neg iqap)$, then no answer is on the record.

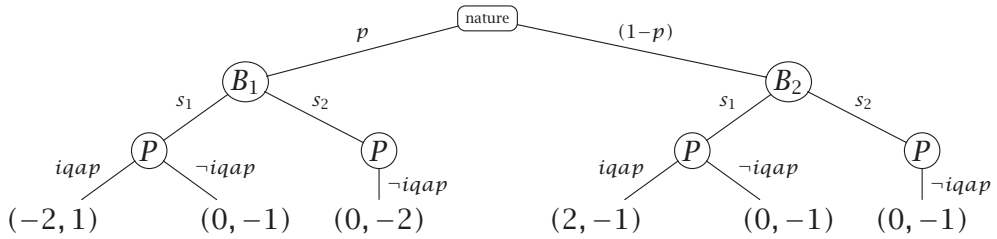


Figure 1 *P*'s model of the game where *B* responds to (1c) and *P* interprets that response. B_1 is a type of Bronston that likes to misdirect *P*, while B_2 prefers not to misdirect *P*. The utilities are in the order (P, B).

P's and *B*'s utilities on the outcomes depend on *B*'s type and on what *both* *P* and *B* do. The particular numbers and even their ratios could have been another way but only up to a point, since their tradeoff against the probability distribution *Pr* determines the choices the players will make. But certainly their *qualitative ordering* matters, and it reflects the following intuitions about *P*'s and *B*'s preferences among the various possible outcomes of the game. First, the optimal outcome for *P* is that *B* is of type B_2 (the version of *B* that doesn't want to misdirect) and they perform $(s_1, iqap)$: this puts an answer on the public record without *B* misdirecting him, as *P* intended *B* to do (an intention he tries to fulfil by asking the question (1c)). However, the utilities for the outcomes for B_2, given that this is *P*'s model of the game and *P* strongly believes *B* to be guilty (regardless of whether he is of type B_1 or B_2), are commensurate with the inevitable risk that B_2 must take. If the outcome is $(s_1, iqap)$, then he will have enhanced his trustworthiness in the eyes of the jury by providing an answer (we will discuss this issue of trust in more detail shortly), but he will also have made himself vulnerable to a perjury conviction (since this outcome puts a false answer on the public record). If the outcome is $(s_2, \neg iqap)$, then he isn't vulnerable to perjury, but he does undermine trust. Since both these outcomes have good points and bad points for B_2, we assign them both the same, slightly negative utility for him. On the other hand, B_1 is penalised more than B_2 in the outcome $(s_2, \neg iqap)$, because not only has B_1 (like B_2) undermined trust in issuing a non-answer, but B_1 also wants to misdirect *P* and this outcome means he failed to do so.

B is also penalised, whatever his type, when P treats an implicature as unsafe (i.e., the outcome is $(s_1, \neg iqap)$): B_1 is penalised because he wants to misdirect P and failed to do so; and B_2 is penalised because he does not want to misdirect P and he inadvertently did! Note that P treating the implicature as unsafe does *not* reveal B to be dishonest and the relatively minor penalty for B_1 reflects this — this is a game where P is simply computing the message, rather than inferring whether the speaker believes that message. We do not consider cases where s_2 is interpreted as *iqap*, since this interpretation of s_2 violates basic cooperativity.[9] Finally, B_1 prefers P to interpret s_1 as an indirect answer: this outcome entails that his preference to misdirect P has been achieved. But P is penalised in this outcome, since he has miscalculated what B committed to (given that B is of type B_1, the version of Bronston that is misdirecting).

As we've argued intuitively, calculating the optimal strategies depends not only on the utility of the individual outcomes but also on the value of p: P's belief of how likely it is that B wants to misdirect him. This probability distribution reflects a different kind of information and a facet of language and linguistic usage that is not directly related to truth conditional content. Following Brown & Levinson (1978), Asher & Quinley (2011) argue that language does not have the role merely to convey or ask for propositional content. Language also affects relationships between speakers and hearers, in particular their reputation and a respect for the autonomy or "distance" of a conversational agent from his interlocutors — his freedom from constraints imposed by them on his possible actions. They show how to model such politeness considerations in a game theoretic framework using the notion of a trust game model (McCabe, Rigdon & Smith 2003). Asher (2012) argues that by adapting Asher and Quinley's trust game model to the situation of questions and their answers, the strategy where a player responds to a question by committing to an indirect answer is in general a preferred strategy, especially in situations where the question-response game may be repeated, as is the case here in our courtroom scenario. Answering a question shows that one takes the interlocutor's question seriously, thus

9 Arguably, a general model of human behaviour supports a (defeasible) inference from B saying *I refuse to answer* to B *believing* a positive answer (in other words, B believes the answer potentially incriminates him and so he gains from not committing to that answer in the trial). But this inferred belief is not a part of the message — there's no public commitment to that belief because it is not a part of the semantic interpretation of the coherence relation *Rejection* that B commits to.

adding to his positive face; and giving more information, even if not directly relevant, increases that positive face. A move that adds to the positive face of the interlocutor also looks good in the eyes of the judge and the jury (Sally (2001) and Mazar & Ariely (2006) echo similar themes by speaking of empathy with the prosecutor or a penchant to stay within social norms). Given these considerations, we will take it to be reasonable for P and for witnesses to the conversation to conclude that B is more likely to be of type B_2 — an agent who prefers *not* to misdirect the prosecutor (or, more quantitatively, we make $p < 0.5$).

With these assumptions about the value of p in place, we can test which combination of strategies in Figure 1 are equilibria. The (pure) equilibria are those combination of pure strategies $\langle \mu^*, \alpha^* \rangle$ that maximise both S's and R's expected utilities, the term "expected" factoring in their respective expectations about what the other player will do (recall that μ^* is a function from types T to signals S, and α^* is a function from signals S to messages M). For S, this means that he won't deviate from μ^*, assuming that R is playing α^*, so long as μ^* satisfies the equation (8) for all $t \in T$: i.e., the expected utility of the signal $\mu^*(t)$ is at least as good as any other signal s that he could send in that state, assuming that R is playing by the strategy α^*.

(8) $\quad U_S(t, \mu^*(t), \alpha^*(\mu^*(t))) = \arg\max_s U_S(t, s, \alpha^*(s))$

Calculating the expected payoff for R of their combined (pure) strategies $\langle \mu^*, \alpha^* \rangle$ is more complicated, because S's speaker type is hidden to R. As we said before, the expected utility of a message m is the average utility of each possible outcome of that message, weighted by the probability that the outcome is achieved. R's actual utility of a particular outcome is dependent on (t, s, m), but R is uncertain about the value of t. Thus, as is traditional in Bayesian models of reasoning with uncertainty, we marginalise out the value of t, given the observed evidence — namely, that S performed s. So the expected utility of the message m that R chooses, given that speaker S said s, is:

$$\sum_{t' \in T} Pr(t'|s) U_R(t', s, m)$$

In words, it is the weighted average of the utility of each possible outcome of R's performing m, given that S performed s, with the weight being R's (probabilistic) posterior belief about S's type, given that R observed S perform s. $Pr(t|s)$ is defined in terms of S's strategy μ^*: by Bayes Rule, $Pr(t|s)$ is

$Pr(s|t)Pr(t)$ (up to a normalising factor); and since we are considering only pure strategies, $Pr(s|t) = 1$ if $\mu^*(t) = s$ and it's 0 otherwise. So the strategy (μ^*, α^*) is optimal for R only if it satisfies equation (9):

$$(9) \quad \sum_{t' \in T \text{ s.t. } \mu^*(t')=s} Pr(t')U_R(t', s, \alpha^*(s)) = \arg\max_m \sum_{t' \in T \text{ s.t. } \mu^*(t')=s} Pr(t')U_R(t', s, m)$$

Overall, then, (μ^*, α^*) is a (pure) NE only if both equations (8) and (9) are satisfied.[10]

With these definitions, we can calculate the expected utilities (EU) of P's responses to B's signal. If $\alpha(s_1) = iqap$, then $EU(\alpha(s_1)) = -2p + 2(1-p) = 2-4p$. If $\alpha(s_1) = \neg iqap$, then $EU(\alpha(s_1)) = 0$. So given our assumptions about the value of p — that is, we have argued that considerations surrounding trust and politeness make $(1-p) \gg p$ — P's optimal response to s_1 is $iqap$ (in other words, he will assume that s_1 makes a negative answer a matter of public record). Clearly, P's only response to s_2 is $\neg iqap$.

Given this, Bronston's optimal tactic if he is of type B_1 (that is, someone who wants to misdirect) is to utter s_1: according to (8) this has an expected utility of 2, compared with his alternative s_2 that yields a utility of 0. If $\mu(B_2) = s_1$, then $EU(\mu(B_2)) = -1$, and if $\mu(B_2) = s_2$, then again $EU(\mu(B_2)) = -1$ (for recall P's optimal responses that we just calculated; $\alpha^*(s_1) = iqap$ and $\alpha^*(s_2) = \neg iqap$). So there are two optimal strategies for B_2, given P's optimal strategies: either s_1 or s_2. Overall, then, we get two (pure) NEs. In both of them, P responds to s_1 with $iqap$ and responds to s_2 with $\neg iqap$ and B_1 says s_1. But in the first one, B_2 says s_1, and in the second B_2 says s_2. So note that this game's equilibria do not (necessarily) reveal B's type if he issues s_1.

Let's reconsider the defeasible inference from (1d) to an answer to (1c). It is based on the assumption of coherence and that Bronston is reacting to the prosecutor's question, which is not plausibly deniable here. We've shown that P's conclusion that B committed to a negative answer is reasonable on game theoretic grounds, using the interpretation game in Figure 1. The game's probability distributions over speaker type and utilities are constrained by factors like negative vs. positive face. But these, we've argued, are also rational. The fact that in the actual dialogue P never did ask the follow

10 This concept solution helps R to decide how to react to S's signal s only if s is not a complete surprise, in that s is in the range of at least one NE strategy μ^*. But we'll ignore situations where S says something completely unexpected; it's not relevant to analysing the type of misdirection we observed in dialogue (1).

up question and moreover attempted subsequently to convict B of perjury suggests that P took the equilibrium move to be as we just described (and in particular, that his optimal way of interpreting (1d) was that it made an answer a matter of public record); so P had a model of B where $Pr(B_2) \gg Pr(B_1)$. On the basis of social factors, we've argued that it is reasonable to have a probability distribution over B's type that satisfies this property.

Nevertheless, though rational, P has arguably made a mistake. P is correct to use the small game to test for the coherence of the response. But given the stakes, he needs to test whether the implicature he has drawn from this small game is plausibly deniable by B. P failed to do the latter adequately, in that he should have — and has failed to — consider signals that could have been but weren't sent and that are intuitively highly relevant. In particular, P overlooked the possibility that B considered uttering a *direct answer* but *chose not to* on the grounds that it was suboptimal for him in some way. Note that in the small game of Figure 1, neither P nor B consider direct answers at all. P's model of the game as it's given in Figure 1 is thus too small, if it aims to isolate those commitments that cannot be defensibly denied, which any adequate test for safety must do. It provides a coherent interpretation and a rational basis for inferring an indirect answer, but it doesn't show that that implicature cannot plausibly be denied. To do the latter, one has to consider signals that *monotonically entail* an answer as opposed to implicating it (as indirect answers do). As Parikh (2001) has argued for other applications, it is at least possible that B considered such a signal and chose not to use it.

In general, any interpreter must isolate a relevant set of signals to consider as part of the game. There are an unbounded number of possible signals that the grammar of the language allows, and there are even an unbounded number of coherent signals in context. The problem of reasoning about interpretation is to choose a set of signals (and their attendant possible messages) that is small enough to effectively perform inference over, but large enough to include all relevant information that might affect inferences about interpretation. The signals of Figure 1 yield an understandable and reasonable result; but it's not one that yields an interpretation that is necessarily part of Bronston's public commitments. To assure the latter, P should have modelled B's and his own decisions using a larger game: specifically, one that includes B's utilities for performing a signal s_3 that *monotonically entails* a negative (or positive) answer to the question (1c). Adding this novel signal expands the range of rhetorical roles that P should consider B is publicly committing to as well. This is because of the constraints on content that the linguistic

conventions of this expanded signal set impose: P must consider not just *iqap* (indirect answerhood) but also *qap* (direct answerhood). Indeed, given that s_3 monotonically entails *qap*, the only action available to R if S performs s_3 is *qap* — any other action would not be a coherent interpretation that abides by basic cooperativity.

More generally to evaluate a misdirection, one must ensure that the equilibrium interpretation survives in a game of sufficient size, as measured by the range of signals that are considered to be a part of the game. This makes our test loosely related to Farrell's (1993) strategy for testing equilibria: that is, ensure what you think is equilibrium play remains so in a big enough game. But since we're modelling safety and not credibility, our conditions on the game being big enough are quite different from Farrell's.[11] Specifically, for a game to be large enough to test reliably for safety, it must include signals that *monotonically entail* the equilibrium public commitment (or message m) that is being tested. Our model makes one crucial assumption concerning expressibility: for any equilibrium play (μ, α) that's being tested, where the linguistic form of the *observed* signal s does *not* monotonically imply ϕ (and so it follows that there is at least one speaker type $t \in T$ such that $\mu(t) = s$), but $\alpha(s) = m$ *does* entail ϕ, there is some signal s' whose linguistic form does monotonically entail the content ϕ. Our claim here, then, is that a game yields a safe inference that the speaker publicly commits to the message m and its implicature ϕ with the signal s in the situation above only if (μ, α) is optimal and the game also contains such signals s'.

Including such signals s' in the game is not only necessary, but sufficient as well. And this means that we can test safety in finite, bounded games, making safe inferences about public commitment computable. To see why including s' suffices, consider the possible preferences in such an extended game. If a speaker (type) who intends to publicly commit to ϕ would prefer the latter signal s' over the original one s, then the original equilibrium (μ, α) won't survive in the game in which the signal s' is taken into consideration. So the speaker's commitment to the implicature ϕ isn't safe. If, on the other hand, the original equilibrium signal s is at least as preferred by all speaker types as the signal s' that monotonically entails ϕ, then this indicates that it

11 For instance, Farrell needs to assume that talk is cheap and there is always a neologism — that is, a signal that wasn't a part of the original game — that expresses X for any subset X of all possible speaker types T. Given our different goals (i.e., to model safety) we can remain agnostic about whether talk is cheap because we don't need to express arbitrary subsets of possible speaker types.

is rational to realise the intention to convey ϕ as an implicature rather than as an entailment. There is no need to compare the equilibrium interpretation $\alpha(s) = m$ that implicates ϕ to any other novel signal: either the novel signal has no interpretation that is consistent with any possible interpretation of the observed signal s, in which case any rational agent who prefers the latter signal s' would never have uttered his actual signal s at all; or the novel signal s' entails content ψ that's distinct from ϕ but is an alternative candidate implicature of the speaker's actual signal s, in which case it will be an appropriate signal for testing the safety of another equilibrium in the game. So in sum, we test the safety of an equilibrium interpretation of the speaker's message by extending the game (if it is not already so extended) to include a signal that monotonically entails the equilibrium interpretation.

With this in mind, we extend P's model of the game in Figure 1 to include a signal s_3 that monotonically entails a negative answer to his question (1c); e.g., s_3 could be the word *no*.[12] As we said earlier, we must likewise extend P's possible responses to include *qap* because of the conventional semantics of the new signal s_3. The extended game is shown in Figure 2. The utilities on the old outcomes are preserved in this extension. Since B_1 would like to misdirect P, s_3 for him is a costly move: the answer is monotonically entailed by s_3 and so attempting to misdirect with this signal defies basic cooperativity. In contrast, B_2 prefers that P accurately infer B's public commitments, and so for B_2 the outcomes $(s_1, iqap)$ and (s_3, qap) are equally good.

Whether the optimal *strategy* has shifted for B, however, will depend on P's optimal reactions to B's moves in this extended game. So let's examine these now. P can respond to s_1 with either *iqap* or $\neg$*iqap*. As before, which of these is optimal depends on whether $p' > 0.5$ (for P is comparing expected utilities $2 - 4p'$ and 0 respectively). Whether P's optimal response to s_1 has shifted or not depends on the probability distributions Pr and Pr' over B's type in the original game and its extension. Thus safety tests for interpreting s_1 as an indirect answer depends on these probability distributions too.

Given the purpose of extending the game — to test for safety — P should preserve as much as possible his model of B's type in the extended game. For instance, P should not change his model of B's (private) beliefs about the dialogue context. More technically, then, we must ensure that whatever

12 In contrast to Parikh's (2001) assumptions about ambiguous vs. unambiguous utterances, inferring that *no* is a direct answer is no more complex in GL than inferring that s_1 is an indirect answer, and so we do not penalise direct answers on the basis of their complexity in interpretation as Parikh would have to.

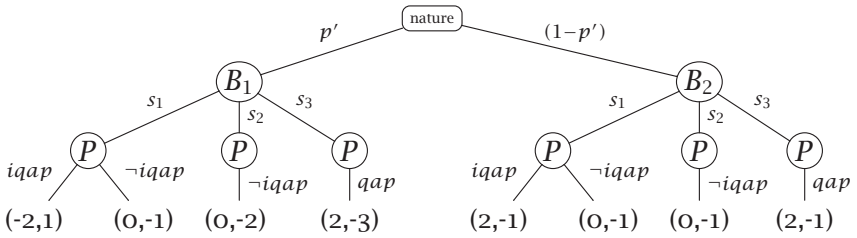


Figure 2 *P*'s model of the *extended* signalling game where *B* responds to P's question (1c) and *P* interprets that response. B_1 is a type of Bronston that likes to misdirect *P*, while B_2 would prefer not to misdirect *P*. The utilities are given in the order (P, B).

P's original likelihood estimates *Pr* were of what type of person *B* is (what *B* believes, whether or not *B* would like to misdirect him etc), the extended model *P* should use a model of *B* that *updates* this *original* model *Pr* of *B* with the fact (or observed evidence) that *B* chose *not* to perform the signal s_3 — in mathematical terms, $Pr(\vec{B}|\neg s_3) = Pr'(\vec{B})$. If the extended game were to adopt a probabilistic model of *B* that deviates from this Bayesian update, then rather than testing the safety of the *original* inference for what's communicated, one would be providing a distinct rationale entirely for inferring what's communicated, based in essence on a completely different model of the speaker, which itself would need to undergo safety tests. Thus testing whether an interpretation of a signal is safe uses extensions that are restricted both in the type of novel moves to be considered (i.e., the only extensions that matter are ones where the novel signal monotonically entails the candidate interpretation) and the probability distribution over speaker types (so that *P* preserves his probabilistic model of *B*). These constraints are captured formally in Definition 2 below, but first we investigate its consequences to our particular example in Figure 2.

We argued earlier that politeness considerations and games of trust justify having a relatively low 'prior' $Pr(B_1) = p$; that is, in the absence of any observed evidence, *P* believes Bronston is unlikely to be an agent who wants to misdirect. But *P*'s probabilistic model *Pr* of *B* should also entail that an agent who wants to misdirect would *virtually never* choose to utter a direct answer — such a move is not a misdirection! On the other hand, *P*'s model of *B* should capture the intuition that were Bronston of type B_2 and intent on putting an answer on public record, then achieving it by uttering s_3

is *guaranteed* to achieve the desired result given the payoffs while uttering s_1 doesn't guarantee it (it depends on whether the interlocutor takes the implicature to be safe). These intuitions about P's model of misdirecting vs. non-misdirecting agents is captured in P's model of B if we make $Pr(B_1|\neg s_3)$ greater than $Pr(B_2|\neg s_3)$: i.e., $p' > 0.5$, in contrast to $p < 0.5$. In other words, while in the absence of any evidence it seems reasonable on the basis of considerations about politeness and trust for P to think B is unlikely to want to misdirect him, the observed evidence — B chose not to supply a direct answer — will increase his likelihood estimate that B's intent is misdirection.

These different prior and posterior beliefs that come into play in the small game and its extension shift the equilibrium. P's optimal strategy α^* is now one where $\alpha^*(s_1) = \neg iqap$ (this has expected utility 0, compared with the alternative response *iqap*, which now has expected utility < 0). $\alpha^*(s_2) = \neg iqap$ as before, and $\alpha(s_3) = qap$, regardless of the value of p'. Because P's optimal strategy shifts, so does Bronston's. The optimal strategy for B_1 is still s_1 (matching the intuitions about the choices that misdirecting agents would make that we just discussed). For note that given P's optimal interpretations α^* in this extended game, the expected utility of $\mu(B_1)$ is -1 if $\mu(B_1) = s_1$ (i.e., $U_B(B_1, s_1, \neg iqap)$); it is -2 if $\mu(B_1) = s_2$ and -3 if $\mu(B_1) = s_3$. The optimal strategy for B_2, on the other hand, has shifted: s_1, s_2 and s_3 are equally optimal for him, all of them having an expected utility of -1. Thus the safety test on P's interpretation of s_1 that we have just performed by extending the game fails: his optimal strategy α^* in the extension is to interpret s_1 as $\neg iqap$. And B issuing s_1 doesn't reveal his type in the extended game (because s_1 is optimal for both types).

We regiment these ideas as follows. Definition 1 defines the type of games that a receiver R constructs when interpreting a speaker S's signal (so in Figure 1, $T = \{B_1, B_2\}$, $S = \{s_1, s_2\}$, and $M = \{iqap, \neg iqap\}$).

Definition 1 **Interpretation Games**

A receiver R's model of interpretation is a game $\langle \mathsf{S}, \mathsf{R}, T, Pr, S, M, U \rangle$ where:

- S is a speaker; R is the receiver;
- T is a set of speaker types;
- $Pr(\vec{T})$ is a probability distribution over T;
- S is a set of signals that S can emit; so a (pure) strategy for S is a function $\mu : T \to S$;

- M is a set of messages or *coherent interpretations* that R can infer; so a (pure) strategy for R is a function $\alpha : S \rightarrow M$; moreover, for any signal $s \in S$ there is at least one message $m \in M$ that is a coherent interpretation of s that is consistent with the compositional and lexical semantics for s.
- $U = \{U_{\mathsf{S}}, U_{\mathsf{R}}\}$ are utility functions from $T \times S \times M$ to $\mathbb{R}$.

The set of extended games that we use to test safety are then formalised as follows:

Definition 2 Game Extension

Let $G = \langle \mathsf{S}, \mathsf{R}, T, Pr, S, M, U \rangle$ be an interpretation game. Then $G' = \langle \mathsf{S}, \mathsf{R}, T', Pr', S', M', U' \rangle$ is a permissible extension of G, written $G' \rhd G$, iff:

i. $T = T'$;

ii. $S \subseteq S'$ and $M \subseteq M'$;

iii. $U \subseteq U'$; that is, for any $(t, s, m) \in T \times S \times M$, $U'_a(t, s, m) = U_a(t, s, m)$ for $a \in \{\mathsf{S}, \mathsf{R}\}$.

iv. Pr' is a Bayesian update of Pr, given the unconsidered signals in $S' \setminus S$. I.e.:

$$Pr'(\vec{T}) = Pr(\vec{T} \mid \wedge_{s_k \in S' \setminus S} \neg s_k)$$

In words, G' is a permissible extension of G if it consists of the same players (and player types), it has strictly more strategies, the utility functions in G' extend those in G, and the probability distribution Pr' is a Bayesian update of the probability function of the original game, given that the moves in S' that were unconsidered originally didn't actually happen.

We can now define when an inference about public commitment is safe. Note that Definition 3 conditions the test for safety on the game being of sufficient size.

Definition 3 Safe Public Commitment

A receiver R's interpretation m of a signal s is a safe representation of a speaker S's public commitment in a game G if and only if in any permissible extension G' of G that includes at

least one signal s' that monotonically entails m, $\alpha^*(s) = m$ remains an optimal strategy for R in the extended game G', according to the equilibria strategies defined by equations (8) and (9).

When applied to (1), the inference to $IQAP(c, d)$ derived from SDRT's glue logic is shown to be unsafe. While it is an optimal interpretation in the small game, it is not in the extended game. The small game *is* the right one for evaluating how *B intended* his signal to be interpreted and how it should be interpreted even in a strategic setting, but it's the wrong one for evaluating whether that interpretation passes the test of plausible deniability. Our model thus implies: (i) the implicature exists and can be rationally drawn in strategic situations, but it is not an unavoidable public commitment of Bronston's, as the prosecutor assumed; (ii) to be safe, the prosecutor should have said something akin to the reporter in (6c). That is, a rational agent would have pursued his goal to get the information about the bank account on the public record with a follow-up, direct question.[13]

5 Cognitive modelling: a symbolic approach

The model in Section 4 leaves room for improvement. It's disconnected from the glue logic (GL); we need to understand how safety interacts with inferences about discourse structure. It's also mute as to the exact nature of the mistake in *P*'s reasoning. Definition 3 stipulates that *P* should consider a signal s_3 that monotonically entails an answer, and so his probability distribution over player types should take such information into account. But it doesn't tell us what led *P* to his mistake in the first place, or how to correct it. What's missing is the *reasoning* that should take us from the smaller game to its extension. The model in Section 4 reflects only the result of reasoning in equilibrium. But intuitively, there is a *dynamic transition* between one model of Bronston's preferences (as he intends to convey them) and another (the better one) that comes from *observing evidence* and from the *reasoning* that takes this new evidence into account. This can't be defined in the framework of the previous section. It can only give the equilibrium strategies in the small game and the bigger one. It doesn't account for the

13 This provides a theoretical reconstruction of the final conclusions by Solan & Tiersma (2005) concerning the Bronston case.

transition from one game to the other, and it is this transition that gives the concept of safety its bite.

To address this gap, we will integrate elements of the model of Section 4 with a logical framework consonant with GL, thus providing an alternative model of strategic conversation. The games in Figures 1 and 2 provide the *model theory* for linking speaker types to public commitments; what we need in addition is a *proof theory* in which the *reasoning* can be explored, in particular the reasoning as to why an agent should consider the larger game rather than the smaller one. So we will start with a partial theory or description of an agent's mental state, which then gets updated and revised as one learns more about the agent or as one considers options that didn't seem relevant before a particular piece of evidence came to light. 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, Gärdenfors & Makinson (1985) suggest. This can either be done by conditionalising a probability distribution over new evidence like the unconsidered move s_3, or, more symbolically, via a theory that incorporates general but defeasible principles about human action and the preferences that underlie them.

To reason about an agent's motives and actions (including discourse actions), we use a symbolic approach and familiar modal operators to express an agent's mental state that integrate easily with our language of nonmonotonic entailment (recall that $A > B$ means *If A then normally B*). We call this language and the axioms developed therein CL (*cognitive logic*). CL as it's developed in Asher & Lascarides 2003 contains modal operators $\mathcal{B}_a$ and I_a: $\mathcal{B}_a\phi$ means agent a believes ϕ, and $I_a\phi$ means agent a intends to bring about a state that entails ϕ. The link between qualitative belief statements $\mathcal{B}_a\phi$ and the probabilistic beliefs as expressed in the games of Section 4 follow straightforwardly, using standard methods for linking quantitative and qualitative belief models (e.g., Pearl (1988)). Finally, our distinction between (private) belief and public commitment leads us to add to CL a modal operator $\mathcal{P}_{a,D}$: the formula $\mathcal{P}_{a,D}\phi$ states that agent a publicly commits to ϕ with regard to the group of agents D.

5.1 Reasoning with partial information about preferences

Our cognitive logic CL needs a way to express and reason about preferences. For this we draw on CP-nets (Boutilier et al. 2004, Domshlak 2002), the formal

details of which are in the Appendix. We use CP-nets because they offer a compact and *qualitative* way of talking about preferences.

The basic ingredient is a *conditional preference*: "if c is true, then I prefer p to q". In our signalling games, the agents have preferences about the values of a signal S (we might take its values to be, say, s_1, s_2 or s_3) and preferences about the values of the messages M (e.g., *iqap*, *qap* and $\neg$*iqap*). An example preference formula is (10), which means: if the signal is s_1, then agent a prefers the message to be *iqap* (i.e., he prefers that the speaker be committed to an indirect answer) rather than $\neg$*iqap* or *qap*, between which he is indifferent.

(10) $s_1 : iqap \succ_a \neg iqap \sim_a qap$

This particular conditional preference statement makes a's preferences over the values for M dependent on those of S. An example formula with no such dependencies is the global preference statement $iqap \succ_a \neg iqap$, which says that a unconditionally prefers the message *iqap* to $\neg$*iqap*.

Given that we work with finite games (see Section 4), our signals S and messages M always have a finite number of values in any interpretation game. We can treat these values $s_1, s_2, \ldots, iqap, \ldots$ as *propositional variables* in our CL, so long as we also include appropriate background axioms: for instance $s_i \wedge (s_i \rightarrow \neg(s_j \vee s_k))$, for $i, j, k \in \{1, 2, 3\}$ and $i \neq j \neq k$ (in words, exactly one signal is emitted). One would need an equivalent set of background axioms for the messages too. Thus it is easy to extend the CL language to express formulae like (10). Note also that an agent may control some of the variables over which he has preferences, and not control others. For those that he controls, one can conceive of the variable as an *action*: it is something that the agent can choose to make true, or not.

As an example, suppose a customer C calls a box office clerk O to book tickets to the opera:

(11) a. C: Do you have two tickets to the opera in the dress circle?

b. O: I can give you two tickets together in the stalls.

O's task in this interpretation game is to respond to C's question (11a) with a signal (here it's (11b)) and then C's task is to compute what content the signal publicly commits O to.

On the basis of the arguments for safe inference given in Section 4, we will assume that the game consists of *three* possible signals and the coherent interpretations that are consistent with their conventional semantics. The

first one s_1 is the signal that O actually performed (i.e., (11b)), which is consistent with the coherent interpretation *iqap* (indirect answerhood) and $\neg$*iqap* (a refusal to answer) but not *qap* (direct answerhood). (11b), like (1d), defeasibly implies a negative answer via shared knowledge (e.g., that the stalls is not a part of the dress circle) on the basis of scalar implicatures. As in Section 4, then, C should include in his model both a signal s_2 that monotonically entails $\neg$*iqap* (e.g., *I refuse to answer*, or *I wonder when my break is*), and a signal s_3 that monotonically entails a negative answer (e.g., *no*, and its attendant message *qap*).[14] The conditional preference statements in (12) then capture intuitively compelling information about C's and O's preferences (recall that O controls the values of the variables s_1, s_2, s_3, and C controls the values of the variables *iqap*, *qap* and $\neg$*iqap*):

(12)

	$iqap \succ_O qap \succ_O \neg iqap$		$iqap \succ_C qap \succ_C \neg iqap$
iqap:	$s_1 \succ_O s_3 \sim_O s_2$	*iqap*:	$s_1 \succ_C s_3 \sim_C s_2$
qap:	$s_3 \succ_O s_1 \sim_O s_2$	*qap*:	$s_3 \succ_C s_1 \sim_C s_2$
$\neg$*iqap*:	$s_2 \succ_O s_1 \succ_O s_3$	$\neg$*iqap*:	$s_2 \succ_C s_1 \succ_C s_3$

At the domain level, both C and O want C to buy tickets to the opera. So both C and O would rather O commit to an answer to C's question than not so commit, and moreover they prefer O to commit to an answer that identifies tickets to buy over an answer that does not. (12) reflects this with the first preference statement for each agent. O and C share *global* preferences over O's public commitments; they both want him to declare a negative answer that also identifies tickets (*iqap*) over just a direct negative answer (*qap*), which is in turn preferred to a refusal to answer ($\neg$*iqap*). Furthermore, both C and O abide by basic cooperativity. (12) reflects basic cooperativity, because the preferred values of signals are dependent on the message and, for example, if the message is *qap*, then the preferred signal for both agents is s_3 while they are indifferent between s_2 and s_1. Interpreting either of the latter two signals as *qap* would violate basic cooperativity.

Since the agents O and C control the values of some of the variables in these preference statements, the statements determine a *game*. To calculate an optimal joint outcome, CP-nets validate two ordering principles when calculating each agent's relative preferences over all the outcomes of the game. The primary principle is that violating more preference statements is

14 In Section 5.2 we show that for games such as this one, where the agents' preferences are aligned, a smaller game that lacks s_3 is big enough to draw safe inferences about public commitments.

worse than violating fewer of them. The secondary principle is that violating a (conditional) preference of something on which your other preferences depend is worse than violating those other preferences. So, for example, (12) yields $s_3 \wedge iqap$ as a preferred outcome over $s_3 \wedge qap$ for both O and C: both of these outcomes violate exactly one preference statement for each agent (the first violates $iqap : s_1 \succ s_3 \sim s_2$ and the second violates $iqap \succ qap \succ \neg iqap$); but violating the global preference statement is worse. In fact, this logic yields for both O and C the same optimal outcome from (12), namely $s_1 \wedge iqap$. Equilibrium play is defined as before: agents act so as to achieve an outcome that no agent would unilaterally deviate from, given the subset of the variables that he controls.

The above principles suffice to decide which outcomes are optimal when complete information about preferences is available. A nice byproduct of this approach, however, is that it applies to situations with only *partial* information about preferences, action and behaviour. For example, suppose Bronston (B) wants the prosecutor (P) to interpret his response to (1c) as an indirect answer, and he prefers to utter signal s_1 (i.e., (1d)) if P would interpret s_1 this way, otherwise he prefers to utter an explicit refusal to answer (signal s_2). In other words, he has the preferences given in (13).

(13)

	$iqap \succ_B \neg iqap$
$iqap$:	$s_1 \succ_B s_2$
$\neg iqap$:	$s_2 \succ_B s_1$

But suppose that B lacks complete knowledge of P's preferences: he does not know if P is a gullible prosecutor, who tends to treat all implicated content as a matter of public record (even if it's not safe to do so; i.e., $s_1 : iqap \succ_P \neg iqap$) or whether he is skeptical and so prefers not to interpret the signal s_1 as a commitment to an answer (i.e., $s_1 : \neg iqap \succ_P iqap$). If P is gullible, then their joint preferences would make B's unique optimal move s_1. If P is skeptical, then preferences would make B's unique optimal move s_2. Consequently, if B doesn't know if P is gullible or skeptical, he will have to use factors that are additional to his partial knowledge of preferences to decide between performing s_1 and s_2.

We handle partial information about preferences in CL thanks to two features. First, CL's language already expresses partial preferences: any strict subset of a complete set of preference statements—e.g., any subset of the statements in (12)—expresses partial information about preferences. Secondly, we exploit *default reasoning* (via the weak conditional $>$) to make

partial information about preferences defeasibly complete, thereby allowing us to use the above two principles for identifying optimal outcomes to decide how to act. In other words, a set of preference statements is made complete by adding assumed preferences that defeasibly follow from default axioms in CL, with the 'last resort' being that agents default to being indifferent among relevant variables for which preference information is missing entirely. So CL will include axioms $A > B$ where B is a conditional preference statement and A expresses information about preferences, public commitment, belief etc. We'll see examples of such axioms shortly.

This approach avoids reasoning about a range of player types, each associated with complete preferences. Instead, when interpreting utterance (1d), say, agent P will reason with and revise his partial description of B's preferences, exploiting the evidence that he said (1d) (see Section 5.2 and 6 for details). Thanks to the nonmonotonic logic of CL's $>$, one can support decisions about what action to perform even if knowledge of preferences is partial.

Let's suppose that G is a game: i.e., a conjunction of conditional preference statements for its two players with each player controlling the value of some subset of the variables. CL must link such a game (or preference statements) G to the agents' *beliefs* when choosing their optimal move: since the optimal outcome for an agent in G can include variables whose values he doesn't control, one needs to check that he doesn't find his optimal state(s) 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 defeasibly follow from the premises and background theory of CL axioms. So to identify an agent's optimal belief-compliant state(s), we filter out from his set of best responses to variables he doesn't control those states that are defeasibly inconsistent with his beliefs (this is decidable). Within CL this leads to the definition of a *CP-solution*$_a(\phi, G)$ for agent a and game G:

Definition 4 *CP-solution*$_a(\phi, G)$ holds iff:

i. a is a player in the game G; and

ii. $s \vdash \phi$ for every belief-compliant optimal state s of G: i.e., where Γ includes the CL background axioms and the relevant premises about players in G, $\Gamma \nvdash \mathcal{B}_a \neg s$ and for

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

For example, if B's preferences are those in (13), and B believes that P's preferences satisfy $s_1 : \neg iqap \succ_P iqap$, then by Definition 4 $\textit{CP-solution}_B(s_2, G)$ holds.

5.2 Principles of action and decision making

With this treatment of preferences in CL we can now provide principles that makes our qualitative and symbolic model approximate the predictions of standard game theoretic models. The principles we provide make agents *pay-off maximisers* (the basic principle of rationality from game theory), *basic cooperative* (Clark 1996, Lewis 1969), and defeasibly *committed to discourse coherence* (the basic principle of SDRT).

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 that 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. $(G \wedge \textit{CP-solution}_a(\psi, G)) > I_a\psi$

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

`Maximising Utility` part (a) ensures a intends ψ if ψ follows from all belief-compliant optimal states (by Definition 4). Indeed, agent a's intentions are conditional not only on *all* of a's beliefs but also *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 4). It yields constraints on G from intentions: if one knows $I_a\psi$ and nothing about G or about a's beliefs, then the minimal preference representation 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 *commitments*, made via 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, which exploits recursion over the conversation's discourse structure, is detailed in Cadilhac et al. (2011), but the details of this aren't relevant for our purposes here.

CL captures *basic cooperativity* via an intention that one's public commitments be shared:

- `Intent to Share Commitment:` $(b \in D \wedge \mathcal{P}_{a,D}\phi) > \mathcal{P}_{a,D} I_a \mathcal{P}_{b,D}\phi$

If a commits to b (among others) to content ϕ, 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 games: why commit to content if you don't intend that others accept the commitment? Following Perrault (1990), we refine this default axiom for assertions: when a commits to an assertion ϕ to b, then normally $\mathcal{P}_{a,D} I_a \mathcal{B}_b \phi$ (this is also stronger than just basic cooperativity).

Finally, we make agents normally commit to a rhetorical connection to some prior contribution: in the axiom `Coherence`, $?_R(?, \phi)$ means that the discourse segment ϕ to which a commits is connected to some available prior segment in the discourse context with some coherence relation:

- `Coherence:` $b \in D \wedge \mathcal{P}_{a,D}\phi > \mathcal{B}_b \mathcal{P}_{a,D} ?_R(?, \phi)$

As with `Intent to Share Commitment`, this assumption is grounded in a game theoretic justification based on face considerations: following Asher & Quinley (2011), `Coherence` is rational on politeness grounds, because completely ignoring all prior dialogue is an affront to the interlocutor's face.

In Section 4, we defined safety as a test involving game extensions. The arguments from Section 4 that adding a signal s' that monotonically entails ϕ is necessary and sufficient for testing the safety of an inference that a

speaker is committed to ϕ still apply. But in this symbolic cognitive model, agents have a single partial theory about preferences, which gets updated and revised through observation: in particular, as new actions are considered, preferences over the *existing* actions may be revised. So adding new actions in CL doesn't always create game extensions, and the test for safety in this symbolic model must therefore be adapted. Specifically, the test for safety now involves extending a *game frame* — a game frame being the set of players and the actions available to each player (in other words, the game frame abstracts away from the players' preference profiles that turn a game frame into a particular game).

Extending a game frame allows inferences in CL to revise preferences for the original actions, in light of the novel actions in the extension. An inference that an implicature ϕ is a part of the speaker's public commitment is safe just in case it remains a part of the equilibrium strategy (i.e., in both agents' CP-solutions) over the game that is defeasibly inferable from the game frame that's extended to include at least one signal s' that entails ϕ. More formally, let S and R be the speaker and receiver in an interpretation game. Let T be R's background theory of general human decision making (e.g., the CL axiom `Maximising Utility`) together with his particular knowledge of S's and his own mental states (including, for instance, preference information extracted from the dialogue context, including the signal that S has just performed). Then T supports defeasible inferences from the set of signals and messages $S \times M$ in their interpretation game to a complete set of preference statements or game $G(S \times M)$ — the 'last resort' being to infer indifference whenever preference information about outcomes in $S \times M$ is missing entirely. We write this defeasible inference as $T + S \times M \mathrel{|\!\sim} G(S \times M)$. Now suppose that an interpretation ϕ is optimal in $G(S \times M)$ — more formally $\textit{CP-solution}_{\mathsf{R}}(\phi, G(S \times M))$. Then ϕ is a safe interpretation just in case the entailment in (14) holds, where $S' \times M'$ is $S \times M$ extended with at least one signal s' (and attendant messages m') such that $s' \vdash \phi$:

(14) $\quad T + (S' \times M') \mathrel{|\!\sim} G(S' \times M') \wedge \textit{CP-solution}_{\mathsf{R}}(\phi, G(S' \times M'))$

This definition of safety emphasises how from the standpoint of the symbolic model, the game frame extensions are just a more specific non-monotonic theory, working over a more specific set of premises (in particular, one that includes additional signals and messages). Safety then amounts to ensuring that an optimal outcome given a small set of observations and considered moves can still be calculated as optimal when considering all

the *relevant* moves — the relevant moves being signals that monotonically entail the optimal message being tested. This is why safety is rationally important: in deciding whether a move is optimal, one should take all the relevant information into account. And it is the (finite) set of coherence relations — in particular, the subset of coherence relations whose semantic consequences are consistent with the compositional and lexical semantics of the observed signal — that specifies the relevant moves in a finite way.

We have now provided all the pieces of our symbolic model, the proof theoretic counterpart to the more traditional model of Section 4. This permits us to provide a correspondence between the Gricean Principles of Section 2 and games where preferences align. We first offer the following definition of Grice Cooperative games:

Definition 5 A game is **Grice Cooperative (GC)** just in case for any of its players a and b

i. When a or b perform a particular speech act, then they normally have the intentions that, according to the glue logic, are associated with a move of that type — so to make an assertion is (normally) to have the intention that the audience believe it; and to ask a question is to (normally) intend to receive a true and informative answer; and

ii. $\mathcal{B}_b(\phi : \psi \succ_a \neg\psi) > (\phi : \psi \succ_b \neg\psi)$
(in other words, the agents believe that their preferences normally align).

Assuming the axioms of `Intent to Share Commitment`, `Maximising Utility` and `Coherence` are mutually believed and several other reasonable axioms about beliefs, intentions and about preferences over beliefs, we can state and prove within CL Fact 1 and Fact 2, which together provide the equivalence of our characterisation of GC environments with the axiomatic approach given in Section 2 (see the Appendix for the proofs).

Fact 1

`Sincerity:` $(\mathcal{P}_{a,D}\phi \wedge GC) > \mathcal{B}_a\phi$

`Sincerity for Intentions:` $(\mathcal{P}_{a,D}\, I_a\phi \wedge GC) > I_a\phi$

`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`: $(b \in D \wedge \mathcal{P}_{a,D} I_a \phi \wedge GC) > I_b \phi$

Fact 1 states principles of Sincerity and Cooperativity that are usually primitive axioms in Belief Desire Intention (BDI) approaches to dialogue; here, we prove that they are *derivable* in CL, given its axioms of rational behaviour, if the dialogue agents D are players in a game G that satisfies Definition 5. They make any *declared* belief, intention or preference in a GC conversation normally an *actual* belief, intention or preference too; `Competence` makes belief transfer the norm; and `Cooperativity` makes a declared individual intention normally a shared actual intention.[15]

We now relativise our Gricean Principles of Sincerity for factual contents, intentions and preferences to an arbitrary interpretation game G (i.e., G may be an uncooperative game). Basically, one simply replaces the formula GC in the principles in Fact 1 with any game G (thereby forming a default with a less specific antecedent). For instance, the relevant form of Sincerity in a game G is (15), and *mutatis mutandis* for the other Gricean Principles.

(15) $(\mathcal{P}_{a,D} \phi \wedge G) > \mathcal{B}_a \phi$

With the Gricean principles so defined, one can prove Fact 2 (the proof is in the Appendix).

Fact 2 Suppose Sincerity, Sincerity for Intentions and for Preferences, Competence and Cooperativity hold of an interpretation game G. Then G is GC.

The Gricean Principles are thus quite strong: by Fact 2, they defeasibly entail alignment of preferences of the two agents, a fact that Lewis' (1969) signalling games assume. Facts 1 and 2 together yield a proof of a folklore fact:

Fact 3 Lewis-style signalling games that verify `Maximising Utility`, `Intent to Share Commitment` and `Coherence` are a special case of GC games and entail all the Gricean Principles.

Intention and belief transfer in a GC conversation is a *default*: even if the preferences align, conflicting prior 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

15 This generalises and strengthens the result of Asher, Sher & Williams (2001) concerning `Sincerity`.

and denial occur in GC dialogues, as exemplified in (16), a GC conversation on the assumption that *A* and *B* share preferences to go to Chop Chop and may even share a preference to there go by car:

(16) 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.

This example also shows that while conversations that are GC (as defined in Definition 5) and conversations that are rhetorically cooperative must both be basic coooperative, GC games and rhetorical cooperativity are independent of one another. (16) is GC according to Definition 5, but it is not rhetorically cooperative because (16b) implicates that *B rejects* the intention revealed by (16a) — to get to Chop Chop by car.

A GC environment yields the more specific CL principles in Fact 1 from CL's general axioms. We likewise have a special result with respect to testing safety in GC conversations. Consider again the GC dialogue (11). Indirect answers are prevalent as responses to questions: SDRT captures this via its GL axiom (5), which makes an indirect answer the default response to any question. So if the compositional semantics of a question α and its response β is consistent with $IQAP(\alpha, \beta)$, then that is an equilibrium in a simple game in which we consider responses broadly, as *iqap* vs. non-answers, or $\neg$*iqap*. A simple game, including just the two signals s_1 (a signal that in GL yields an inference to *iqap*) and s_2 (a signal that does not yield a GL inference *iqap*) is given in (17):

(17)

	$iqap \succ_O \neg iqap$		$iqap \succ_C \neg iqap$
iqap:	$s_1 \succ_O s_2$	*iqap*:	$s_1 \succ_C s_2$
$\neg$*iqap*:	$s_2 \succ_O s_1$	$\neg$*iqap*:	$s_2 \succ_C s_1$

By the earlier discussion (see equation (14)), testing the safety of this focal equilibrium involves adding a novel signal s_3 that *monotonically*, rather than defeasibly, entails the answer (e.g., s_3 is *no*). So one likewise adds *qap* to the game too. The result of this is already given in (12). This extension retains the same CP-solution, namely (s_1, *iqap*). The game in (12) reflects how, in GC environments, indirect answers may be preferred to direct ones because of the extra information they provide. Interpreting (11b) as an indirect answer is safe because relative to the agents' underlying game problem it is a preferred response over simply providing the direct answer *no*: it anticipates the customer's CP-solution on learning the direct answer and helps him to

achieve that. More generally, given that in GC environments the defaults needed to infer scalar and other implicatures are sound, *iqap* moves, and indeed all discourse computations of coherence relations and hence equilibria of our 'smaller' interpretation games, are normally safe. Using the principles in Fact 1, safe interpretations are normally credible too:[16]

Fact 4 A CP-solution of a GC interpretation game is normally safe and credible.

Thus, strategic reasoning in a GC environment can be carried out with small games: their equilibrium results will normally hold in the relevant extensions. Thus strong cooperativity simplifies strategic thinking.

6 Non cooperative conversations

The GC condition doesn't hold in the courtroom dialogue (1). So the principles in Fact 1 and Fact 4 don't apply, though `Maximising Utility`, `Intent to Share Commitment` and `Coherence` do.

Interpreting (1d) involves using defeasible inference to construct *P*'s model of *B*'s and his own preferences over the signals that *B* contemplates as responses to (1c), and *P*'s interpretation of those signals, with *P*'s defeasible inference about preferences taking observations about *B*'s chosen action (i.e., the signal (1d)) into account. Let's assume that whatever method one uses for extracting commitments to preferences from signals yields the intuitive result from *B*'s actual signal s_1 (or (1d)): i.e., that *B* prefers *iqap* (i.e., *B* prefers *P* to interpret him as committed to a negative answer) over $\neg$*iqap* — thus matching the predictions about logical form from the glue logic GL. Then the definition of safety in CL demands that we reason with a preference description that includes not only s_1 (and *iqap*) but also a signal s_3 that *monotonically* entails the negative answer (e.g., *no*). In other words, we must work with an *extended game frame*, so that in testing safety *P* essentially compares two games: the set of strategies in one game (consisting of the actual signal s_1) is a subset of the other (consisting of signal s_3 too), but the games may *disagree* on the preferences over those strategies. That's because

16 To illustrate this with our example game, *O* uses the Gricean principles to interpret *C*'s response as implicating a negative answer (see Section 2), and by `Sincerity` and `Competence` *C* believes this answer. This is a simple, symbolic counterpart to the much more elaborate result in Crawford & Sobel 1982.

the information in the larger game leads one to revise one's estimates of the other agents' preferences.

As in Section 4, we illustrate the comparison with a smaller game containing just the signals s_1 (the string (1d)) and s_2 (*I refuse to answer*) and the messages *iqap* and $\neg$*iqap*. *P*'s model of his own and *B*'s preferences is shown in (18):

(18)

	$iqap \succ_B \neg iqap$		$iqap \succ_P \neg iqap$
$iqap$:	$s_1 \succ_B s_2$	s_1:	$iqap \succ_P \neg iqap$
$\neg iqap$:	$s_2 \succ_B s_1$	s_2:	$\neg iqap \succ_P iqap$
		$iqap$:	$s_1 \succ_P s_2$
		$\neg iqap$:	$s_1 \sim_P s_2$

As in Figure 1, *P* would prefer *B* to commit to an answer (i.e., *iqap*), but *only if* the signal's conventional semantics is consistent with this commitment (so his preferences over interpretations are dependent on the signal). Indeed, if his interpretation is $\neg$*iqap*, then he's indifferent about the signal (see *P*'s utilities in Figure 1). In contrast to the GC games in (12) and (17), *P*'s preferences feature a dependency of the message on the signal and *vice versa*. *P*'s model of *B* assumes that *B* wants *P* to interpret his signal as an answer: after all, this is what *B* would want, whether he wants to misdirect or not. But *P* also reasons that if *B* thinks *P* is the skeptical prosecutor described earlier (i.e., a prosecutor who performs $\neg$*iqap*), then *B* will prefer to refuse to answer (so *B*'s preferences over signals depend on how *P* will interpret it).

The logic of preferences described earlier yields from the preference statements in (18) the following (partial) preference orderings over all outcomes for *B* and *P*:

(19) $(s_1 \wedge iqap) \succ_B (s_2 \wedge iqap) \succ_B (s_2 \wedge \neg iqap) \succ_B (s_1 \wedge \neg iqap)$
$(s_1 \wedge iqap) \succ_P \{(s_1 \wedge \neg iqap), (s_2 \wedge \neg iqap)\} \succ_P (s_2 \wedge iqap)$

In this small game, $(s_1 \wedge iqap)$ is an optimal solution for both players; it is irrational for Bronston to choose a signal other than s_1 and for the prosecutor to choose an interpretation of s_1 other than *iqap*.

But this inference isn't safe. It constitutes a defeasible deduction about *B*'s and *P*'s preferences given the 'small' game frame consisting of signals s_1 and s_2 — a deduction that is reasonable and normally sound given general principles of human behaviour as well as particular facts about the situation at hand. But it ignores potentially highly relevant information. The definition of safety in CL (see (14)) demands that *P* compute defeasible inferences about preferences over a game frame that includes a signal s_3 that monotonically

entails the implicated answer that is a part of the equilibrium strategy ($s_1 \wedge$ *iqap*) in the smaller game. The preferences over the game with the extended strategy set — where s_3 is added to S and *qap* is added to M — is shown in (20).

(20)

$$iqap \succ_B \neg iqap \succ_B qap \qquad s_3 \succ_P s_1 \sim_P s_2$$

iqap:	$s_1 \succ_B s_2 \succ_B s_3$	s_3:	$qap \succ_P iqap \sim_P \neg iqap$
$\neg$*iqap*:	$s_2 \succ_B s_1 \succ_B s_3$	s_1:	$\neg iqap \succ_P iqap \sim_P qap$
qap:	$s_3 \succ_B s_1 \sim_B s_2$	s_2:	$\neg iqap \succ_P iqap \sim_P qap$

Observe how the preferences change: the preference $s_1 : iqap \succ_P \neg iqap$ is reversed in the larger game. Indeed, P's preferences over the signal S are now global.

These changes occur as a consequence of P's (defeasible) reasoning about preferences over the more specific scenario, which includes s_3 as well as s_2 and s_1. He must, in particular, reason as to why B considered s_3 but chose not to perform it. Intuitively, he concludes (defeasibly) that B believes that were he to utter a signal s_3 whose clear interpretation is *qap*, he would either have to lie explicitly thereby risking perjury or to admit to a damaging piece of evidence. B prefers *iqap* and even $\neg$*iqap* over *qap*. This counterfactual reasoning isn't representable at all in the orthodox model of Section 4, but it can be represented in our qualitative model CL, which supports generic reasoning over mental states and human action (though we don't give specific formal details of this counterfactual reasoning here). With P's different model of his own and B's preferences, which are defeasibly inferred for the extended game frame, P's optimal interpretation of B's move s_1 is $\neg$*iqap* rather than *iqap*. Furthermore, by reasoning about preferences in the extended game frame P infers why it is rational for B to utter s_1: namely, B prefers that P perform a defeasible but unsafe inference about what B has publicly committed to (for note that B's optimal outcome remains (s_1, *iqap*) even in the bigger game, but that is no longer a CP-solution for P).

From our proof theoretic perspective then, safety encodes the passage from a "general" context in which certain reasonable but defeasible inferences are drawn to a more specific context in which different perhaps incompatible inferences are drawn. The prosecutor's mistake in the Bronston trial was similar to someone who given a default theory with the principles *birds fly*, *penguins don't fly* and *penguins are birds* and an observation that B is a bird concludes that B flies. This inference is unsafe because he should check first whether B is a penguin! The analogous test with interpretation

games is to check whether it is rational for an agent to exploit implicatures derivable in the general context to misdirect. If so, then these implicatures may be deniable; i.e., the interpreter cannot assume that such implicatures are an undeniable part of the public record. We have argued this test can be achieved within finite games by extending the game to include a signal (and its attendant coherent interpretation) that monotonically entail the implicature.

Our model of safety predicts that unlike (1d), interpreting (1d′) as an indirect answer is safe.

(1) d′. Bronston: Only my company had an account, for about six months, in Zurich.

(1d′) must be connected to a prior segment to make an antecedent *alternatives set* available for *only*. As before, the salient interpretation among the possible coherent alternatives is $IQAP(c, d')$; we test its safety by extending the game frame with a direct answer s_3. In contrast to (1d), interpreting (1d′) as *iqap* does *not* rely on a scalar implicature borne from Bronston providing all relevant information to yield the negative answer. Instead, it simply relies on an assumption that (1d′) is attached to (1c): this resolves the alternatives set to $\{bronston, company\}$, thereby making (1d′) imply a negative answer thanks to the lexical semantics of *only*. But the signal s_3 relies on exactly the same assumption — that it attach to (1c) — to entail an answer. So (s_3, *qap*) cannot be strictly dispreferred by B to (1d′, *iqap*). That is, in contrast to (1d), including s_3 in the game does not compel P to revise his model of B's preferences over *iqap* and hence his own preferences too: the default inference that (1d′) commits Bronston to an answer is therefore safe.

Our model also accounts for cases where coherence isn't preserved like (6). Sheehan's responses cannot connect to the questions with *IQAP*. While the antecedent to the GL axiom (5) is satisfied, its consequent is inconsistent with compositional semantics. Instead, coherent interpretations of Sheehan's response (6b) that are consistent with its compositional semantics include one that provides *Commentary* on it and one where it's not rhetorically connected to (6a) at all. Both of these implicate a refusal to answer. This implicature is safe because it is preferred to the novel move *I refuse to answer* thanks to politeness considerations (it is less confrontational to implicate the rejection of another agent's goal than to express it explicitly).

7 Related work

Our theory differs from Gricean accounts in two fundamental ways. First we have derived Gricean cooperativity from a more general game-theoretic axiomatisation of human behaviour that also handles non-cooperative conversation. We have replaced Gricean formalisations that adopt plan recognition techniques or shared intentions (e.g., Grosz & Sidner (1990), Allen & Litman (1987), Lochbaum (1998)) with a formalisation in terms of shared preferences, thereby allowing rejection and denial to be content-cooperative when the agents' conflicting beliefs yield conflicting optimal solutions (or, equivalently, intentions) for achieving their shared preferences. In addition, we have replaced Gricean derivations of implicatures based on shared intentions with a different picture in which discourse coherence and its attendant implicatures emerge as a rational preference.

Game theory offers rational reconstructions of cooperative conversational implicatures (e.g., Parikh (2001), van Rooij (2004), Benz, Jäger & van Rooij (2005), Asher, Sher & Williams (2001)). But instead of treating cooperativity as a starting point, our model makes it a special case. Our approach refines and extends the Grounding Acts Model of dialogue (Traum 1994) by providing a logical underpinning to its update rules (Poesio & Traum 1997): the update rules articulate particular effects on content and on cognitive states of various dialogue actions; our model makes such effects *derivable* and moreover provides a basis for articulating additional update rules for non-cooperative moves.

Our work draws from the literature on signalling games (e.g., Crawford & Sobel 1982, Farrell 1993, Lipman 2003, Lipman & Seppi 1995). We bring to this work a more sophisticated mapping of signals to meaning: the overly simple signalling models used in prior work don't distinguish between what is literally said from what is implicated, and therefore cannot make locutionary content safe and implied content unsafe, which we argued is crucial to understanding examples like (1). Prior signalling models also fail to distinguish credible inference from safe inference, because either the signal is taken to mean whatever it is optimal for it to mean, or, as in Farell's (1993) model, the mapping from signal to meaning is completely fixed, which ignores the complex interplay of pragmatics, semantics and strategic reasoning.

Pinker, Nowak & Lee (2008) use game theory to predict when plausible deniability is optimal, making implicating content preferable to explicitly expressing it. But their linguistic model does not use constraints on interpre-

tation from discourse coherence, and so they don't distinguish the plausible deniability of the implied answer in (1d) vs. (1d′). We also showed how exploiting coherence avoids the need for testing an unlimited number of equilibria for safety, thanks to the finite number of coherence relations in the ontology (e.g., in SDRT there are around 30 relations) and the finite ways in which the semantics of such relations can fix the underspecified aspects of meaning revealed by linguistic form (e.g., the finite number of antecedents that they make available to anaphora). This paper has not addressed, however, how one might evaluate the *credibility* of a message: in other words, we have not provided a model that predicts when a (safe) interpretation of a speaker's signal matches his beliefs about the world, and when it doesn't.

Franke, de Jager & van Rooij (2009) analyse non-cooperative conversation by adapting the principle of optimal relevance from Relevance Theory (Sperber & Wilson 1995) within classical game theory. Their account is incomplete, with no formal details of the adapted principle. Providing the formal details will be problematic, we think, because inferences in classical game theory rely on complete information about preferences and this isn't always plausible in conversation. A similar observation applies to the game-theoretic approach to political debate in Klebanov & Beigman 2010 and to work on *dialogue games* (e.g., Amgoud 2003, MacKenzie 1979, McBurney & Parsons 2002, Walton & Krabbe 1995, Wooldridge 2000). Models of decisions in partially observable environments have been developed (e.g., Roth, Simmons & Veloso (2006)). But while these endow agents with uncertainty about the current state and the outcomes of actions, they still rely on full knowledge of the *possible* states and preferences over them, which isn't always realistic for dialogue agents. In contrast to these approaches, the utilities for each possible dialogue move in our CL need not be pre-defined or quantified. Skyrms (1994, 2004) models the adoption of linguistic conventions in the absence of complete information about preferences within an evolutionary framework. We are working at a different level of analysis, however, because we are interested not only in basic dispositions to linguistic behaviour but also how this behaviour may change in the light of information brought to bear just in the current conversation and through reasoning.

Franke (2010) takes an *epistemic approach* to pragmatic interpretation. He adopts a slightly different game-theoretic solution concept, known as Iterated Best Response or IBR. But the more important differences concern the underlying conception of the game. He uses a signalling game in which the conventional meaning $[\![s]\!]_{\mathcal{L}}$ of the speaker's message s is a subset of

speaker types and hence denotes the states of affairs in which s is true. But he does not consider aspects of linguistic meaning that have to do with coherence as we do, and so he does not distinguish between the meaning of the signal based only on lexical and compositional semantics and what we have called the message m whose content incorporates elements derived from inferences about coherence. For Franke, $[\![s]\!]_{\mathcal{L}} = [\![m]\!]_{\mathcal{M}}$ and both are fixed and monotonic. Implicatures arise as *beliefs* because IBR predicts an optimal interpretation that is different from $[\![s]\!]_{\mathcal{L}}$: for example, it may be more specific than $[\![s]\!]_{\mathcal{L}}$ (i.e., $\alpha^*(s) \subset [\![s]\!]_{\mathcal{L}}$; e.g., scalar implicatures); or not (i.e., $\alpha^*(s) \not\subseteq [\![s]\!]_{\mathcal{L}}$, making the message incredible and the speaker insincere). In our model, the receiver's action is to determine what the public commitment of the speaker is, independently of whether he finds the message credible or not. Because inferences about coherence can depend on strategic factors, modelling safety becomes necessary. One must test one's inferences about what $[\![m]\!]_{\mathcal{M}}$ is, because its value is computed under uncertainty with partial information (like all inferences about coherence are). Consequently, the linguistic meaning of the message $[\![.]\!]_{\mathcal{M}}$ can and does get revised as agents reason about the game. More generally, while we acknowledge that some pragamtic inferences are epistemic in nature (e.g., see `Sincerity`), we take pragmatic inferences that occur through establishing coherence to be a matter of public commitment, and not simply a matter of updating one's model of the private mental attitudes of the speaker, as Franke does. It is only because we treat certain pragmatic inferences as surfacing in public commitments that we can distinguish between perjury convictions for (1d) (interpreting it as putting an answer on the public record isn't safe) vs. (1d′) (interpreting it as putting an answer on the public record is defeasible, but safe).

Constraints imposed by discourse coherence, which we believe is a key and primitive component of any good model of conversation, limit the size of the game; coherence, allows us to estimate what actions are irrelevant and therefore can be excluded from reasoning. This reduces the complexity problem of computing an optimal and safe interpretation of a signal and makes it workable, even when a player has partial information about the interpretation game that he's playing.

8 Conclusions

We have argued that any model of strategic conversation must distinguish between the public commitments one makes through utterances and private mental states that affect and are affected by them; it must make discourse coherence a key and primitive principle for reasoning about the speaker's public commitments; and it must acknowledge that computing the speaker's public commitments is often based on defeasible inference. It must also allow for several levels of cooperativity; strategic conversation typically exhibits only some of them. We focused on cooperativity and lack thereof at the level of *interpretation*.

We introduced the concept of safety, which provides a sufficient condition for what content of an utterance is a matter of public record. We distinguished this notion from the game theoretic notion of credibility, which tests whether the content of a message can be believed. Safe implicatures must be part of public commitments, while those that are not may be defensibly denied. The inference to an indirect answer from the signal (1d) is not safe, and though the implicature is reasonable (it is optimal in the small interpretation game), we have shown that Bronston could defensibly have denied that he was committed to the implicature.

We proposed two alternative game-theoretic models for evaluating safety. Both models involve making the game of sufficient size (in terms of the set of signals that the speaker is taken to be contemplating). Both models crucially rely on discourse coherence: coherence determines when a game is of sufficient size to make interpretations safe, but not so large that computing optimal interpretations becomes unworkable. But the two models are also quite different. The first model from Section 4 uses standard tools of non-cooperative games. The second (see Sections 5 and 6) is a symbolic qualitative model that makes safe interpretation a byproduct of sound nonmonotonic inference generally: safety is the strategic counterpart to the *methodological principle* that one must consider *all* relevant evidence before drawing any conclusion based on defeasible principles. In essence, the symbolic model offers a correspondence result between a nonmonotonic theory for computing optimal moves and a game theoretic model.

We showed the qualitative model in contexts where the agents share preferences to be equivalent to the Gricean Principles of Section 2. This shows that games which are subject to the Gricean Principles are a strict subset of all interpretation games. It also allowed us to show that in such

Gricean environments, we can use smaller strategic reasoning games to compute interpretations that are optimal, safe and credible. So Gricean cooperativity is in some sense more efficient.

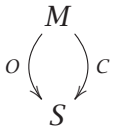
While we have a correspondence theorem for Gricean Cooperative games, we don't have one for non cooperative conversational games. The naturally occurring examples that motivated our model exhibit only a tiny fraction of the full range of strategic moves. But the lack of available empirical data means no one even knows what those are. Our framework thus invites extensive empirical work on strategic conversations, something we plan to do.

Appendix

CP-nets. Our proofs of Facts 1 and 2 rely on the fact that the preference statements we described in Section 5 are in fact partial descriptions of **CP-nets** (Boutilier et al. 2004, Domshlak 2002). So we start with a brief overview of CP-nets, focussing on the aspects of their language and logic relevant to our proofs.

A CP-net has two components: a directed *conditional preference graph* (CPG), which defines for each variable V its set of parent variables $Pa(V)$ that affect the agent's preferences over V's values; and a *conditional preference table* (CPT), which specifies for each variable V the agent's preferences over V's values, given every combination of values in $Pa(V)$. A CP-net for a *game* consists of a CP-net for each player.

For example, the CPG corresponding to the example CPT given in (12) is the following:



To maximise compactness, we have assumed S takes three values s_1, s_2, s_3 (and similarly M takes three values) rather than showing the CPG over the six boolean variables themselves.

While dependencies among variables can be read off the CPT, many theorems concerning equilibria strategies and the complexity in computing them are dependent on the properties of the CPG; e.g., whether it contains cycles (e.g., see Apt, Rossi & Venable 2005 for details). Our proofs of Fact 2 will likewise exploit the configuration of the CPG (i.e., whether or not it contains

cycles) and indeed will proceed by induction on the CPG.

As we said in Section 5.1, the logic of CP-nets follows the following two ranked principles when generating the preference order over every combination of actions 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 variable is worse than violating the preference on a daughter variable. Research on CP-nets has yielded highly efficient methods that capture at least some *Nash Equilibria* (NE).[17] CP-nets also have a lot of expressive power: Apt, Rossi & Venable (2005) show that there is a translation of games into CP-nets, where a game has an NE a just in case a is an optimal solution of its CP-net.

Outline Proof of Fact 1

`Sincerity:` $(\mathcal{P}_{a,D}\phi \wedge GC) > \mathcal{B}_a\phi$
Assume that all the normal GC consequences hold; we show that the conclusion must hold. Accordingly, suppose $\mathcal{P}_{a,D}\phi$ and GC hold. We assume also that ϕ expresses a proposition that is capable of being believed. By `Intent to Share Commitment`, it defeasibly follows that $\mathcal{P}_{a,D}\, I_a\mathcal{B}_b\phi$ for any $b \in D$. By the first part of GC, the coherence relations that one normally infers in GL and their associated speech act related goals (or SARGs) are assumed to hold. So $I_a\mathcal{B}_b\phi$ for any $b \in D$. By `Maximising Utility` and the fact that I is an S4 modality, $I_a\mathcal{B}_b\phi$ defeasibly implies $\mathcal{B}_b\phi \succ_a \neg\mathcal{B}_b\phi$. By the second clause of GC, $\mathcal{B}_b\phi \succ_b \neg\mathcal{B}_b\phi$.

We assume that preferences over an agent's beliefs pattern after her factual preferences:

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

So $\phi \succ_b \neg\phi$. By GC again, we have $\phi \succ_a \neg\phi$. By our assumption (1), we have

(2) $\mathcal{B}_a\phi \succ_a \neg\mathcal{B}_a\phi$.

We claim that a should normally believe ϕ in all the normal worlds picked out by our assumptions and axioms, if he prefers in reflective equilibrium to have the belief states he in fact has. For $\neg\mathcal{B}_a\phi$ would imply $\neg\mathcal{B}_a\phi \succ_a \mathcal{B}_a\phi$, which contradicts (2). So we have defeasibly deduced $\mathcal{B}_a\phi$ from the premises (plus the assumption that (1) is a valid axiom of the logic). Now, `Weak Deduction` is a valid rule of the weak conditional $>$ (Asher 1995): if $\Gamma, \phi \mid\!\sim \psi$, $\Gamma \not\mid\!\sim \psi$ and

17 See for instance Bonzon 2007 for results on acyclic nets.

$\Gamma \nvdash \neg(\phi > \psi)$ then $\Gamma \vdash (\phi > \psi)$. Conditionalising on the reasoning from the assumptions to $\mathcal{B}_a\phi$, we have $(\mathcal{P}_{a,D}\phi \wedge GC) \vdash \mathcal{B}_a\phi$. □.

`Sincerity for Intentions:` $(\mathcal{P}_{a,D} I_a\phi \wedge GC) > I_a\phi$
Suppose $\mathcal{P}_{a,D} I_a\phi \wedge GC$. By `Sincerity` (which we've just proved), $\mathcal{B}_a I_a\phi$. Assuming that intentions are doxastically transparent (i.e., $\mathcal{B}_a I_a\phi \rightarrow I_a\phi$) yields the result 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:` $(\mathcal{P}_{a,D}\phi \wedge \mathcal{P}_{b,D}?\phi \wedge a, b \in D) \rightarrow ((\mathcal{B}_b\mathcal{B}_a\phi \wedge GC) > \mathcal{B}_b\phi)$
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 we are in a GC environment. Given the first condition in Definition 5 on GC games, the intention that normally underlies asking a question (to know an answer) and `Maximising Utility` ensures that b's asking a question 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 an equilibrium 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, b must believe a to be trustworthy, and so $\mathcal{B}_b\phi$. Using `Weak Deduction` thus yields `Competence`. □.

`Cooperativity:` $(b \in D \wedge \mathcal{P}_{a,D} I_a\phi \wedge GC) > I_b\phi$
Assume $b \in D \wedge \mathcal{P}_{a,D} I_a\phi \wedge GC$. By `Sincerity for Intentions`, we have $I_a\phi$. 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 $I_b\phi$. Using `Weak Deduction` gets us the desired $>$ statement. □.

Outline Proof of Fact 2

We need to show two things for G to be GC: 1) that in G, players normally have the intentions that are conventionally associated with the speech acts they perform; and 2) that preferences normally align. Proving 1) is relatively straightforward, assuming that each contribution is made with certain con-

versational goals in mind that are known by all conversational participants. Thus, if a player asks a question, the commonly known default conversational goal associated with that speech act is that the speaker intends to have a true and informative answer to the question. By cooperativity of intentions, it follows that the player issuing a response to the question will adopt such a goal, and by rationality will try to realise that goal with an appropriate speech act — e.g., answering the question or saying that he has not enough information to answer it. Similar reasoning applies to other speech acts and conversational goals.

Part 2 is more involved. We prove it by induction on the complexity of the preferences of agents — i.e., by induction over the complexity of CP-nets. What we prove is this: that CP-nets over publicly committed preferences are isomorphic as long as the CP-nets are *realistic* in the following sense: $\psi\colon \phi \succ_n \neg\phi \rightarrow \mathcal{B}_n(\psi \rightarrow \Diamond\phi)$. In words: if player n prefers ϕ over $\neg\phi$ in a context ψ, then he believes that ϕ is *consistent* with ψ being true. Because the preferences are public commitments, we can take them to be mutually believed given Sincerity and Competence (for a proof of this fact, see Asher & Lascarides 2008).

Lemma 1 If the Gricean Principles hold in a game G, and the players' CP-nets are acyclic and realistic, then normally they are isomorphic.

If a CP-net is acyclic, the inductive rank of a given CP-net is obvious, and the proof is straightforward. We assume both players have a simplest CP-net with just one variable and it contains just one preference statement without any dependencies, which in Boolean terms is of the form $\psi \succ_n \neg\psi$ for player n. Suppose $\mathcal{B}_2(\psi \succ_1 \neg\psi)$. Then by Mutual Belief, $\mathcal{B}_1\mathcal{B}_2(\psi \succ_1 \neg\psi)$. By Competence, $\mathcal{B}_1(\psi \succ_1 \neg\psi)$. Since preferences are doxastically transparent, $\psi \succ_1 \neg\psi$. Since the CP-net is realistic, ψ is belief compliant, i.e., $\mathcal{B}_1\Diamond\psi$. By Mutual Belief, $\mathcal{B}_2\mathcal{B}_1\Diamond\psi$ and so $\mathcal{B}_2\Diamond\psi$, by Competence. Since ψ is belief compliant, in the CP-net for 1, ψ is a CP-net solution, and so by `Maximise Utility` $I_1\psi$. By Cooperativity of Intentions, $I_2\psi$. By `Maximise Utility` and the fact that ψ is belief compliant for 2, ψ is the CP-net solution for 2. Since the CP-net has only 1 variable, $\psi \succ_2 \neg\psi$. A similar situation holds for player 2's commitments to preferences.

Now for the inductive step. Suppose that players 1 and 2 agree on preferences in their CP-nets up to rank k. Now consider an acyclic CP-net of rank $k+1$ with a new variable ψ, which depends on some Boolean combi-

nation of values ϕ and ϕ' in a subnet of rank $i < k$. Assume $\mathcal{B}_2(\phi\colon \psi \succ_1 \neg\psi \wedge \phi'\colon \neg\psi \succ_1 \psi)$. By the induction hypothesis player 1 and 2's preferences align over ϕ and ϕ', and then our prior reasoning for the base step completes the proof of Lemma 1.

Lemma 2 If the players in a game G adopt the Gricean Principles and their CP-nets have a simple cycle and are realistic, normally their CP-nets are isomorphic.

If the CP-nets involved have cycles, the proof is more involved because cycles render problematic the rank of a CP-net. But if we consider a simple cyclic net with two variables, we can get alignment of preferences in either of two ways. Consider a game with 2 players again, where 1 has without loss of generality, $\phi\colon \psi \succ_1 \neg\psi$, $\neg\phi\colon \neg\psi \succ \psi$, and player 2 has $\psi : \phi \succ_2 \neg\phi$ and $\neg\psi\colon \neg\phi \succ_2 \phi$, where 1 controls ψ and 2 controls ϕ. These preference statements mean that we have a dependency of ϕ on ψ and ψ on ϕ. We have two CP-net solutions: $\phi \wedge \psi$ and $\neg\psi \wedge \neg\phi$. Using the reasoning of the acyclic case for each agent, we can get them to the point:

$$I_1\phi : \psi \succ_1 \neg\psi \wedge I_1\neg\phi : \neg\psi \succ_1 \psi$$
$$I_2\psi : \phi \succ_2 \neg\phi \wedge I_2\neg\psi : \neg\phi \succ_2 \phi$$

By Cooperativity of Intentions, then, they mutually adopt the other's intentions so that we have

$$I_{1,2}\phi : \psi \succ_1 \neg\psi \wedge I_{1,2}\neg\phi : \neg\psi \succ_1 \psi$$
$$I_{1,2}\psi : \phi \succ_2 \neg\phi \wedge I_{1,2}\neg\psi : \neg\phi \succ_2 \phi$$

This shows that they have aligned preferences.

Finally, we can decompose every CP-net into a set of cyclically dependent variables, the associated tables of which are acyclic. By treating each cycle as a unit, we can now turn the CP-net into an acyclic one. By doing induction on the size of the cycles using Lemma 2 and on the rank of the transformed CP-net using Lemma 1, the general result follows. □

References

Alchourrón, Carlos, Peter Gärdenfors & David Makinson. 1985. On the logic of theory change: partial meeting contraction and revision functions. *Journal of Symbolic Logic* 50. 510–530. http://dx.doi.org/10.2307/2274239.

Allen, James & Diane Litman. 1987. A plan recognition model for subdialogues in conversations. *Cognitive Science* 11(2). 163–200. http://dx.doi.org/10.1207/s15516709cog1102_4.

Amgoud, Leila. 2003. A formal framework for handling conflicting desires. *Proceedings of ECSQARU 2003*. http://dx.doi.org/10.1007/978-3-540-45062-7_45.

Apt, Krzysztof, Francesca Rossi & Kristen Venable. 2005. CP-nets and nash equlibria. *Proceedings of the Third International Conference on Computational Intelligence, Robotics and Autonomous Systems*.

Asher, Nicholas. 1993. *Reference to abstract objects in discourse*. Kluwer Academic Publishers. http://dx.doi.org/10.1007/978-94-011-1715-9.

Asher, Nicholas. 1995. Commonsense entailment: a conditional logic for some generics. In Gabriela Crocco, Luis Farinas & Andreas Herzig (eds.), *Conditionals: from philosophy to computer science*, 103–145. Oxford University Press.

Asher, Nicholas. 2012. The non cooperative basis of implicatures. *Proceedings of Logical Aspects of Computational Linguistics 2012*. http://dx.doi.org/10.1007/978-3-642-31262-5_3.

Asher, Nicholas. 2013. Implicatures and discourse structure. *Lingua* 132. 13–28. http://dx.doi.org/10.1016/j.lingua.2012.10.001.

Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Asher, Nicholas & Alex Lascarides. 2008. Commitments, beliefs and intentions in dialogue. *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (Londial)*. 35–42. http://www.kcl.ac.uk/innovation/groups/ds/events/assets/paperasher.pdf.

Asher, Nicholas & Jason Quinley. 2011. Begging questions, their answers and basic cooperativity. *Proceedings of the 8th International Conference on Logic and Engineering of Natural Language Semantics (LENLS)*. http://dx.doi.org/10.1007/978-3-642-32090-3_2.

Asher, Nicholas, Itai Sher & Madison Williams. 2001. Game theoretic foundations for pragmatic defaults. *Amsterdam Formal Semantics Colloquium*.

Benz, Anja, Gerhard Jäger & Robert van Rooij (eds.). 2005. *Game theory and pragmatics*. Palgrave Macmillan. http://dx.doi.org/10.1057/9780230285897.

Bonzon, Elise. 2007. *Modélisation des interactions entre agents rationnels: les jeux booléens*. Université Paul Sabatier, Toulouse PhD thesis. http://tel.archives-ouvertes.fr/docs/00/23/92/94/ANNEX/soutenance.pdf.

Boutilier, Craig, Ronen Brafman, Carmel Domshlak, Holger Hoos & David Poole. 2004. CP-nets: a tool for representing and reasoning with conditional *ceteris paribus* preference statements. *Journal of Artificial Intelligence Research* 21. 135–191. http://dx.doi.org/10.1613/jair.1234.

Brown, Penelope & Stephen Levinson. 1978. *Politeness: some universals and language usage*. Cambridge University Press.

Cadilhac, Anais, Nicholas Asher, Farah Benamara & Alex Lascarides. 2011. Commitments to preferences in dialogue. *Proceedings of the 12th Annual SIGDIAL Meeting on Discourse and Dialogue*. 204–215. http://www.aclweb.org/anthology/W11-2023.

Clark, Herb. 1996. *Using language*. Cambridge, England: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511620539.

Crawford, Vincent & Joel Sobel. 1982. Strategic information transmission. *Econometrica* 50(6). 1431–1451. http://dx.doi.org/10.2307/1913390.

Domshlak, Carmel. 2002. *Modeling and reasoning about preferences with CP nets*. Ben Gurion University PhD thesis.

Farrell, Joseph. 1993. Meaning and credibility in cheap-talk games. *Games and Economic Behaviour* 5. 514–531. http://dx.doi.org/10.1006/game.1993.1029.

Franke, Michael. 2010. Semantic meaning and pragmatic inference in non-cooperative conversation. In Thomas Icard & Reinhard Muskens (eds.), *Interfaces: Explorations in logic, language and computation* (Lecture Notes in Artificial Intelligence), 13–24. Berlin, Heidelberg: Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-14729-6_2.

Franke, Michael, Tikitu de Jager & Robert van Rooij. 2009. Relevance in cooperation and conflict. *Journal of Logic and Language*. http://dx.doi.org/10.1093/logcom/exp070.

Geurts, Bart. 1996. Local satisfaction guaranteed. *Linguistics and Philosophy* 19. 259–294. http://dx.doi.org/10.1007/BF00628201.

Ginzburg, Jonathan. 2012. *The interactive stance: meaning for conversation*. Oxford University Press.

Green, Mitchell. 1995. Quantity, volubility and some varieties of discourse. *Linguistics and Philosophy* 18(1). 83–112. http://dx.doi.org/10.1007/BF00984962.

Green, Mitchell. 2000. Illocutionary force and semantic content. *Linguistics and Philosophy* 23(5). 435–473. http://dx.doi.org/10.1023/A:1005642421177.

Grice, H. Paul. 1975. Logic and conversation. In Peter Cole & Jerry Morgan (eds.), *Syntax and semantics volume 3: speech acts*, 41–58. Academic Press.

Grosz, Barbara & Candice Sidner. 1990. Plans for discourse. In Jerry Morgan Philip Cohen & Martha Pollack (eds.), *Intentions in communication*, 417–444. MIT Press.

Hamblin, Charles. 1987. *Imperatives*. Blackwells.

Hobbs, Jerry. 1979. Coherence and coreference. *Cognitive Science* 3(1). 67–90. http://dx.doi.org/10.1207/s15516709cog0301_4.

Hobbs, Jerry, Martin Stickel, Douglas Appelt & Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence* 63(1–2). 69–142. http://dx.doi.org/10.1016/0004-3702(93)90015-4.

Kehler, Andrew. 2002. *Coherence, reference and the theory of grammar*. CSLI Publications, Cambridge University Press.

Klebanov, Beata & Eyal Beigman. 2010. A game-theoretic model of metaphorical bargaining. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 698–709. http://www.aclweb.org/anthology/P10-1072.

Lascarides, Alex & Nicholas Asher. 2008. Agreement and disputes in dialogue. *Proceedings of the 9th SigDial Workshop on Discourse and Dialogue (SIGDIAL)*. 29–36.

Lewis, David. 1969. *Convention: a philosophical study*. Harvard University Press.

Lipman, Barton. 2003. Language and economics. In Marcelo Basili, Nicola Dimitri & Itzchak Gilboa (eds.), *Cognitive processes and rationality in economics*. London: Routledge.

Lipman, Barton & Duane Seppi. 1995. Robust inference in communication games with partial provability. *Journal of Economic Theory* 66. 370–405. http://dx.doi.org/10.1006/jeth.1995.1046.

Lochbaum, Karen. 1998. A collaborative planning model of intentional structure. *Computational Linguistics* 24(4). 525–572.

MacKenzie, James. 1979. Question begging in non-cumulative systems. *Journal of Philosophical Logic* 8. 117–233. http://dx.doi.org/10.1007/BF00258422.

Mann, William & Sandra Thompson. 1987. Rhetorical structure theory: a framework for the analysis of texts. *International Pragmatics Association Papers in Pragmatics* 1. 79–105.

Mazar, Nina & Dan Ariely. 2006. Dishonesty in everyday life and its policy implications. *Journal of Public Policy and Marketing* 25(1). 1–21. http://dx.doi.org/10.1509/jppm.25.1.117.

McBurney, Peter & Simon Parsons. 2002. Dialogue games in multi-agent systems. *Informal Logic. Special Issue on Applications of Argumentation in Computer Science* 22(3). 257–274.

McCabe, Kevin, Mary Rigdon & Vernon Smith. 2003. Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization* 52(2). 267–275. http://dx.doi.org/10.1016/S0167-2681(03)00003-9.

Osborne, Martin & Ariel Rubinstein. 1990. *Bargaining and markets*. Academic Press.

Parikh, Prashant. 2001. *The use of language*. CSLI Publications.

Pearl, Judea. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kauffmann.

Perrault, Ray. 1990. An application of default logic to speech act theory. In Jerry Morgan Philip Cohen & Martha Pollack (eds.), *Intentions in communication*, 161–186. MIT Press.

Pinker, Stephen, Martin Nowak & James Lee. 2008. The logic of indirect speech. *Proceedings of the National Academy of Science* 105(3). http://dx.doi.org/10.1073/pnas.0707192105.

Poesio, Massimo & David Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence* 13(3). 309–347. http://dx.doi.org/10.1111/0824-7935.00042.

van Rooij, Robert. 2004. Signalling games select Horn strategies. *Linguistics and Philosophy* 27. 493–527. http://dx.doi.org/10.1023/B:LING.0000024403.88733.3f.

Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1(1). 75–116. http://dx.doi.org/10.1007/BF02342617.

Roth, Maayan, Reid Simmons & Manuela Veloso. 2006. What to communicate? execution time decision in multi-agent POMDPs. *Proceedings of the International Symposium on Distributed Autonomous Robotic Systems (DARS)*. http://dx.doi.org/10.1007/4-431-35881-1_18.

Sally, David. 2001. On sympathy and games. *Journal of Economic Behaviour and Organization* 44(1). 1–30. http://dx.doi.org/10.1016/S0167-2681(00)00153-0.

van der Sandt, Rob. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics* 9(4). 333–377. http://dx.doi.org/10.1093/jos/9.4.333.

Schulz, Katrin. 2007. *Minimal models in semantics and pragmatics: free choice, exhaustivity, and conditionals*. University of Amsterdam PhD thesis.
Skyrms, Brian. 1994. *Evolution of the social contract*. Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511806308.
Skyrms, Brian. 2004. *The stag hunt and the evoluation of social structure*. Cambridge University Press.
Solan, Lawrence & Peter Tiersma. 2005. *Speaking of crime: the language of criminal justice*. Chicago, IL: University of Chicago Press. http://dx.doi.org/10.1111/j.1540-5893.2006.00278_9.x.
Sperber, Dan & Deirdre Wilson. 1995. *Relevance: communication and cognition*. Blackwells.
Stalnaker, Robert. 2002. Common ground. *Linguistics and Philosophy* 25(5). 701–721. http://dx.doi.org/10.1023/A:1020867916902.
Traum, David. 1994. *A computational theory of grounding in natural language conversation*. Computer Science Department, University of Rochester PhD thesis.
Traum, David, William Swartout, Jonathan Gratch & Stacy Marsella. 2008. A virtual human dialogue model for non-team interaction. In Laila Dybkjær, Wolfgang Minker & Nancy Ide (eds.), *Recent trends in discourse and dialogue*, vol. 39 (Text, Speech and Language Technology), 45–67. Springer Netherlands. http://dx.doi.org/10.1007/978-1-4020-6821-8_3.
Walton, Douglas & Erik Krabbe. 1995. *Commitment in dialogue*. SUNY Press.
Wooldridge, Michael. 2000. *Reasoning about rational agents*. MIT Press.

Nicholas Asher
IRIT, Université Paul Sabatier
118 route de Narbonne
F-31062 Toulouse Cedex 4
France
asher@irit.fr

Alex Lascarides
School of Informatics
University of Edinburgh
10, Crichton Street
Edinburgh, EH8 9AB
United Kingdom
alex@inf.ed.ac.uk