Commitments, Beliefs and Intentions in Dialogue

Nicholas Asher
IRIT
Université Paul Sabatier, Toulouse
asher@irit.fr

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

Abstract

We define grounding in terms of shared public commitments, and link public commitments to other, private, attitudes within a decidable dynamic logic for computing implicatures and predicting an agent's next dialogue move.

1 Introduction

A theory of dialogue should link discourse interpretation to general principles of rationality and cooperativity (Grice, 1975). The so-called 'mentalist approach' treats dialogue as a function of the agents' attitudes, usually formalised with BDI (belief, desire, intention) logics (e.g., Grosz and Sidner (1990)). Grounding a proposition p—by which we mean that all dialogue agents mutually agree that p is true—occurs when the BDI logic implies that p is mutually believed.

However, there are compelling reasons to reject the mentalist approach to dialogue modelling. Gaudou et al. (2006) use (1) to argue for a distinction between grounding and mutual belief.

(1) a. A to B (C out of earshot): C is stupid.
b. B to A (C out of earshot): I agree.
c. A to B (C in earshot): C is smart.

(1a) is grounded for A and B. If B now utters *That's right*, then (1c) should be grounded for A and B too. So if grounding is a function of mutual belief, then A and B would hold contradictory beliefs, making them irrational. But A is not irrational; he is disingenuous. Gaudou et al. (2006) conclude that grounding is a function of *shared public commitments*, following Hamblin (1987). But the link to other attitudes is also essential: B should detect that A is lying—i.e., that he can't believe everything that he has publicly committed to.

Dialogue (1) contrasts with dialogue (2), where A 'drops' a commitment to (2a) in favour of (2b), making (2b) grounded:

(2) a. A: It's raining.
b. B: No it's not.
c. A: Oh, you're right.

A theory of dialogue should distinguish between A's illocutionary act in (1c) vs. (2c), even though in both cases A asserts the negation of his prior assertion.

In this paper, we propose a framework for dialogue analysis that synthesises Hamblin's commitment-based approach with the mentalist approach. We think both perspectives on dialogue are needed. In Lascarides and Asher (2008), we argue that the commitment-based framework captures facts about grounding, making explicit the distinction between what is said and private attitudes. But the BDI view is essential for strategic reasoning about dialogue moves. We draw on the strengths of both approaches while avoiding some of their weaknesses. For instance, we avoid the uncomputable models of discourse that stem from default reasoning in first-order BDI logics.

Our starting point is SDRT (Asher and Lascarides, 2003). In Section 2 we modify its representation of dialogue content so that it tracks the public commitments of each dialogue agent. In Section 3 we reconstruct its separate, but related, cognitive logic (CL) to include the attitude of public commitment and axioms that relate it to other, private, attitudes. CL will be a dynamic logic of public announcement, extended with default axioms of rationality and cooperativity. The result will capture the sort of prac-

Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue, June 2–4, 2008, London, U.K.

Turn	A's SDRS	B's SDRS
1	$\pi_1 : K_{\pi_1}$	$\emptyset$
2	$\pi_1 : K_{\pi_1}$	$\pi_{2B} : \mathit{Explanation}(\pi_1, \pi_2)$

Table 1: The logical form of dialogue (3).

tical reasoning that goes on in conversation, when agents adjust their beliefs, preferences and intentions in light of what's said so far. This refines the approach to dialogue using dialogue games (e.g., Amgoud (2003)) because the utilities for each possible dialogue move need not be 'pre-defined' or quantified. Rather, CL will exploit the dynamics in the logic to infer qualitative statements about the *relative* utility of different moves. Furthermore, by approximating game-theoretic principles within the logic, we also deepen the theory by *deriving* some of the cognitive axioms of rationality and cooperativity from them: for instance, a general axiom of Cooperativty (that B normally intends what A intends) will be validated this way. Our approach can also be viewed as extending the Grounding Acts Model (Traum, 1994), providing its update rules with a logical rationale for constraining the update effects on content vs. cognitive states.

2 Dialogue Content

Lascarides and Asher (2008) argue that relational speech acts or *rhetorical relations* (e.g., *Narration*, *Explanation*) are a crucial ingredient in a model of grounding. One of the main motivations is implicit grounding: representing the illocutionary contribution of an agent's utterance via rhetorical relations reflects his commitments to another agent's commitments, even when this is linguistically implicit. For example, B's utterance (3b) commits him to (3a) because the relational speech act $\mathit{Explanation}(3a, 3b)$ that he has performed entails (3a):

(3) a. A: Max fell.
b. B: John pushed him.

Accordingly, the commitments of an individual agent are expressed as a Segmented Discourse Representation (SDRS, Asher and Lascarides (2003)): this is a hierarchically structured set of labelled contents, as shown in each cell of Table 1—the logical form for dialogue (3). For simplicity, we have omitted the representations of the clauses (3a) and (3b) (labelled π_1 and π_2 respectively), and we often gloss the content labelled by π as K_π, and mark the root label of the speaker i's SDRS for turn j as π_{ji}.

The logical form of dialogue is the logical form of each of its turns (where a turn boundary occurs whenever the speaker changes). The logical form of each turn is a set of SDRSs, one for each dialogue participant. Each SDRS represents *all* the content that the relevant agent is currently publicly committed to, from the beginning of the dialogue up to the end of that turn (see Lascarides and Asher (2008) for motivation). And each agent constructs the SDRSs for all other agents, as well as his own—e.g., A and B both build Table 1 for dialogue (3).

The logical form of dialogue (2) is Table 2. Recognising that B's utterance π_2 attaches to π_1 with *Correction* is based on default axioms in SDRT's *glue logic*—i.e., the logic for constructing logical form (Asher and Lascarides, 2003). The content of (2c) (labelled π_3) supports a glue-logic inference that π_3 acknowledges π_2. This resolves π_3's underspecified content to entail K_{π_2}, and so $\mathit{Correction}(\pi_1, \pi_3)$ is also inferred, as shown. In contrast, the fact that (1c) is designed to be overheard by C while (1ab) is not forces a glue-logic inference that they are not rhetorically linked at all; see the logical form in Table 3.

The dynamic semantics for Dialogue SDRSs (DSDRSs) is defined in terms of SDRSs: a DSDRS consists of an SDRS for each participant at each turn, and accordingly the semantics of a dialogue turn is the product of the dynamic semantics for each constituent SDRS. Lascarides and Asher (2008) define grounding at a given turn as the content that's entailed by each SDRS for that turn. Given that each turn represents *all* an agent's 'current' public commitments, the interpretation of a dialogue overall is that of its last turn. Table 2 receives a consistent interpretation, but Table 3 is inconsistent because A's final SDRS is inconsistent. The DSDRS of (3) makes (3a) grounded and that for (2) makes (2b) grounded. The DSDRS of (1) makes (1a) grounded, and should B acknowledge (1c), then anything is grounded.

3 Cognitive Modelling

With this background concerning dialogue content in place, we turn to the interaction of commitments with other attitudes. SDRT's cognitive logic (CL)

Turn	A's SDRS	B's SDRS
1	$\pi_1 : K_{\pi_1}$	$\emptyset$
2	$\pi_1 : K_{\pi_1}$	$\pi_{2B} : \mathit{Correction}(\pi_1, \pi_2)$
3	$\pi_{3A} : \mathit{Correction}(\pi_1, \pi_3) \wedge \mathit{Acknowledgement}(\pi_2, \pi_3)$	$\pi_{2B} : \mathit{Correction}(\pi_1, \pi_2)$

Table 2: The logical form of dialogue (2).

Turn	A's SDRS	B's SDRS
1	$\pi_1 : K_{\pi_1}$	$\emptyset$
2	$\pi_1 : K_{\pi_1}$	$\pi_{2B} : \mathit{Acknowledgement}(\pi_1, \pi_2)$
3	$\pi_{3A} : K_{\pi_1} \wedge K_{\pi_3}$	$\pi_{2B} : \mathit{Acknowledgement}(\pi_1, \pi_2)$

Table 3: The logical form of (1).

supports reasoning about agents' cognitive states in virtue of what they say. Since it contributes directly to constructing the logical form of dialogue, its complexity must be decidable: Asher and Lascarides (2003, p78) argue that this is necessary to explain why, as Grice (1975) claims, people by and large agree on what was said (if not on whether it's true). CL must support default reasoning and hence consistency tests, since agents never have complete information about the dialogue context. And so SDRT makes its CL decidable by denying it access to a dialogue's full, dynamic interpretation—for instance, existentially-quantified SDRS-formulae lose their structure when transferred into CL, thereby losing the relationship between, say, the SDRS-formulae $\neg\exists x\neg\phi$ and $\forall x\phi$.

SDRT's CL from Asher and Lascarides (2003) is deficient in at least two ways. First, it does not support the logical forms from Section 2; CL should include public commitment and its links to other attitudes. Secondly, CL is static, thereby failing to show how attitudes change during dialogue. To overcome these deficiencies we exploit a dynamic logic of public announcement (Baltag et al., 1999). We extend it to support *default* reasoning from public announcements, including (default) inferences about cognitive states. A model $\mathcal{M}$ of the logic consists of a set of worlds $W^{\mathcal{M}}$ and a valuation function $V^{\mathcal{M}}$ for interpreting the non-logical constants at $w \in W^{\mathcal{M}}$. We write $[\![\phi]\!]^{\mathcal{M}} =_{def} \{w \in W^{\mathcal{M}} : \mathcal{M}, w \models \phi\}$. Public announcements are dynamic in that they change the input model into a different output one: any worlds from the input model which fail to satisfy the monotonic consequences of the announcement are eliminated from the output model; likewise for *ceteris paribus* announcements, any worlds that fail to satisfy the nonmonotonic consequences of the announcement are eliminated. More formally, monotonic consequences of an announcement are expressed by the formula $[!\phi]\psi$, where $[!\phi]$ is a modal operator (in words, ψ follows from announcing ϕ). Nonmonotonic consequences are expressed as $[!\phi]^{cp}\psi$, which in turn is defined via a modal connective: $\phi > \psi$ means that *If ϕ then normally ψ*. The model $\mathcal{M}$ therefore also includes a function $*$ from worlds and propositions to propositions, which defines normality and is used to interpret $\phi > \psi$:

$$\mathcal{M}, w \models \phi > \psi \text{ iff } *^{\mathcal{M}}(w, [\![\phi]\!]^{\mathcal{M}}) \subseteq [\![\psi]\!]^{\mathcal{M}},$$

In words, ψ is true in all worlds where, according to w, ϕ is normal. The above description of how announcements transform input models is then formalised in Figure 1.

$$\begin{array}{rl}
\mathcal{M}, w \models [!\phi]\psi \text{ iff} & \mathcal{M}^{\phi}, w \models \psi \\
\mathcal{M}, w \models [!\phi]^{cp}\psi \text{ iff} & \mathcal{M}^{cp(\phi)}, w \models \psi \\
\text{where} & \\
\mathcal{M}^{\phi} = & \langle W^{\phi}, *^{\mathcal{M}}|W^{\phi}, V|W^{\phi}\rangle \text{ where} \\
W^{\phi} = & [\![\phi]\!]^{\mathcal{M}} \\
\mathcal{M}^{cp(\phi)} = & \langle W^{cp(\phi)}, *^{\mathcal{M}}|W^{cp(\phi)}, V|W^{cp(\phi)}\rangle \text{ where} \\
W^{cp(\phi)} = & \{w' \in W^{\mathcal{M}} : \\
& \quad \mathrm{Th}(\mathcal{M}), \phi \mathrel{|\!\sim} \psi \rightarrow \mathcal{M}^{\phi}, w' \models \psi\}
\end{array}$$

Figure 1: Model transitions for announcements

To ensure that CL reflects the commitments in DS-DRSs, we assume that agents announce to the dialogue participants certain commitments to SDRS-formulae. Actually, given the way we have set things

up, each turn commits a speaker to commitments from earlier turns, unless he disavows one of those commitments. $\mathcal{P}_{a,D}\psi$ means that a publicly commits to group D to ψ. Thus a speaker a uttering K_π to D will result in CL-based reasoning with the modality $[!\mathcal{P}_{a,D}\phi_\pi]^{cp}$, where ϕ_π is the shallow representation of K_π (i.e., without existentials). We make the modality $\mathcal{P}_{a,D}$ K45 (one commits to all the consequences of one's commitments, and one has total introspection on commitments, or lack of them), and we also add axioms Ax1 (a commitment to D is a commitment to all its subgroups), and Ax2 (there is a group commitment by x and y to D iff x and y both make that commitment to D):

K: $\mathcal{P}_{a,D}(\phi \rightarrow \psi) \rightarrow (\mathcal{P}_{a,D}\phi \rightarrow \mathcal{P}_{a,D}\psi)$
4: $\mathcal{P}_{a,D}\phi \rightarrow \mathcal{P}_{a,D}\mathcal{P}_{a,D}\phi$
5: $\neg\mathcal{P}_{a,D}\phi \rightarrow \mathcal{P}_{a,D}\neg\mathcal{P}_{a,D}\phi$
Ax1: For any $D' \subseteq D$, $\mathcal{P}_{a,D}\phi \rightarrow \mathcal{P}_{a,D'}\phi$
Ax2: $\mathcal{P}_{\{x,y\},D}\phi \leftrightarrow (\mathcal{P}_{x,D}\phi \wedge \mathcal{P}_{y,D}\phi)$

So the models $\mathcal{M}$ have suitably constrained accessibility relations $R^{\mathcal{P}_{a,D}} \subseteq W \times W$ for all a and D.

Since commitment lacks axiom D, $\mathcal{P}_{a,D}(p \wedge \neg p)$ is satisfiable, reflecting A's public commitments in (1). This contrasts with the belief modality $\mathcal{B}_a\phi$, which is KD45 (with a transitive, euclidean and *serial* accessibility relation $R^{\mathcal{B}_a}$ in the model).

Agent a announcing something to group D will bring about in CL a transition on models: the input model will be updated by adding to a's commitments to D. Changing a model by adding ϕ to a's commitments is defined in equation (4): this stipulates that one adds ϕ to the accessibility relation $R^{\mathcal{P}_{a,D}}_{\mathcal{M}}$, so long as doing so is consistent. Equation (5) defines a similar model transition for beliefs; we'll use this shortly to represent Sincerity.

(4) $\mathcal{M} \mapsto \mathcal{M}_{\phi,a,D}: R^{\mathcal{P}_{a,D}}_{\mathcal{M}_{\phi,a,D}} = (?\top; R^{\mathcal{P}_{a,D}}_{\mathcal{M}}; ?\phi)$

(5) $\mathcal{M} \mapsto \mathcal{M}_{\flat_a\phi}: R^{\mathcal{B}_a}_{\mathcal{M}_{\flat_a\phi}} = (?\top; R^{\mathcal{B}_a}_{\mathcal{M}}; ?\phi)$

We can now interpret announcements about commitments. In words, should an agent a say ϕ to D, then the model is udpated so that all non-monotonic consequences of a's commitment to ϕ are satisfied (so long as this update is consistent):

- `Announcements of Commitment`:
  $\mathcal{M}, w \models [!\mathcal{P}_{a,D}\phi]^{cp}\psi$ iff $\mathcal{M}^{cp(\phi)}_{\phi,a,D}, w \models \psi$

In fact, we assume that should a say K_π to D, then in CL the *ceteris paribus* consequences of this announcement include a's commitment to all glue-logic inferences χ about the illocutionary effects of K_π (as represented via rhetorical relations in the DSDRSs): i.e., $[!\mathcal{P}_{a,D}\phi_\pi]^{cp}\mathcal{P}_{a,D}\chi$. This yields $[!\mathcal{P}_{B,\{A,B\}}\phi_{\pi_2}]^{cp}\mathcal{P}_{A,\{A,B\}}$*Explanation*$(\pi_1, \pi_2)$ in CL from dialogue (3), for instance. Thus the outcome in CL is a model that satisfies $\mathcal{P}_{B,\{A,B\}}$*Explanation*(π_1, π_2), and so long as enough of the semantics of *Explanation* is transferred into CL, this entails (by axiom K) $\mathcal{P}_{B,\{A,B\}}\phi_{\pi_1}$, where ϕ_{π_1} is the shallow representation (3a). A's announcement (3a) ensures the CL model also satisfies $\mathcal{P}_{A,\{A,B\}}\phi_{\pi_1}$. So the CL model reflects what's grounded according to the DSDRS. Table 2, the representation of dialogue (2), yields a CL model that satisfies $\mathcal{P}_{\{A,B\},\{A,B\}}\phi_{\pi_2}$ and $\mathcal{P}_{\{A,B\},\{A,B\}}\neg\phi_{\pi_1}$, where ϕ_{π_1} and ϕ_{π_2} represent (2a) and (2b) respectively. And Table 1 yields a CL model where $\mathcal{P}_{A,\{A,B\}}(p \wedge \neg p)$, p being the (shallow) CL representation of (1a).

An agent's beliefs must be updated at least defeasibly on discovering his commitments. The following `Sincerity` axiom ensures this, by default:

- `Sincerity`: $\mathcal{P}_{a,D}\phi > \flat_a\phi$

We have stated `Sincerity` dynamically via the action operator $\flat_a$; this is the action of updating beliefs and has the following semantics:

- `Belief Update`:
  $\mathcal{M}, w \models \flat_a\phi$ iff $\mathcal{M}_{\flat_a\phi}, w \models \mathcal{B}_a\phi$

`Sincerity` is a default because of examples like (1). As we saw earlier, `Announcements of Commitment` yields $\mathcal{P}_{A,\{A,B\}}(p \wedge \neg p)$. This satisfies the antecedent to `Sincerity`, but $\mathcal{B}_A(p \wedge \neg p)$ is not inferred because it's inconsistent. $\mathcal{P}_{A,\{A,B\}}p$ and $\mathcal{P}_{A,\{A,B\}}\neg p$ are also true (by axiom K); they both satisfy the antecedent of `Sincerity`, but their consequences $\mathcal{B}_A p$ and $\mathcal{B}_A\neg p$ are mutually inconsistent, and so neither is inferred. Thus B detects from A's inconsistent current commitments that he's lying, and without further information B does not know what A believes: p, $\neg p$ or neither one. C, on the other hand, who knows only $\mathcal{P}_{A,\{A,B,C\}}\neg p$, uses `Sincerity` to infer $\mathcal{B}_A\neg p$.

As is standard, mutual belief ($MB_{x,y}\phi$) is defined in terms of belief using a fixed point equation:

(6) $MB_{x,y}\phi \leftrightarrow (\mathcal{B}_x(\phi \wedge MB_{x,y}\phi) \wedge \mathcal{B}_y(\phi \wedge MB_{x,y}\phi))$

This definition means $MB_{x,y}\phi$ entails an ω sequence of nested belief statements: $\mathcal{B}_x\phi, \mathcal{B}_y\mathcal{B}_x\phi, \ldots$ and $\mathcal{B}_y\phi, \mathcal{B}_x\mathcal{B}_y\phi, \ldots$. We will denote a formula that starts with $\mathcal{B}_x$, and alternates with $\mathcal{B}_y$ to a nesting of depth n as $\mathcal{B}^n_{(x,y)}\phi$; similarly for $\mathcal{B}^n_{(y,x)}\phi$. Then one can prove the following scheme is sound.

- `Induction Scheme`:
  Assume $\Gamma \mathrel{|\!\sim} \mathcal{B}_y(\phi \wedge \mathcal{B}_x\phi) \wedge \mathcal{B}_x(\phi \wedge \mathcal{B}_y\phi)$
  And for any n, $\frac{\Gamma \mathrel{|\!\sim} \mathcal{B}_y(\phi \wedge \mathcal{B}^n_{(x,y)}\phi) \wedge \mathcal{B}_x(\phi \wedge \mathcal{B}^n_{(y,x)}\phi)}{\Gamma \mathrel{|\!\sim} \mathcal{B}_y(\phi \wedge \mathcal{B}^{n+1}_{(x,y)}\phi) \wedge \mathcal{B}_x(\phi \wedge \mathcal{B}^{n+1}_{(y,x)}\phi)}$
  Then: $\Gamma \mathrel{|\!\sim} MB_{x,y}\phi$

These axioms ensure that, as in the BDI account, grounding and mutual belief are linked; but unlike the BDI account they are *not* equivalent. Where $D = \{x, y\}$, the **proof** that $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} MB_{x,y}\phi$ is as follows:

1. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_x\phi$ — Ax2, Sincerity
2. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_y\phi$ — Ax2, Sincerity
3. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_y\mathcal{B}_x\phi$ — 1; CL is mutually believed
4. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_y(\phi \wedge \mathcal{B}_x\phi)$ — 2, 3; $\mathcal{B}$ is KD45
5. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_x\mathcal{B}_y\phi$ — 2; CL is mutually believed
6. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} \mathcal{B}_x(\phi \wedge \mathcal{B}_y\phi)$ — 1, 5; $\mathcal{B}$ is KD45
7. $\mathcal{P}_{\{x,y\},D}\phi \mathrel{|\!\sim} MB_{x,y}\phi$ — 4,6; Induction Scheme

□

Thus grounded content is normally mutually believed; e.g., it is in (2) and (3), but not in (1).

Announcements affect intentions as well as beliefs. For instance, an intuitively compelling axiom is `Intent to Ground`: if a commits to ϕ, then normally he commits that he intends (written $\mathcal{I}_a$) that his interlocutors commit to it too, if they haven't done so already. A version of `Sincerity` also applies to intentions, and like `Sincerity` for beliefs requires adding an action operator $\sharp_a$ with a similar interpretation to $\flat_a$, to effect a model transition for the update of intentions.

- `Intent to Ground`:
  $(b \in D \wedge \mathcal{P}_{a,D}\phi \wedge \neg\mathcal{P}_{b,D}\phi) > \mathcal{P}_{a,D}\mathcal{I}_a\mathcal{P}_{b,D}\phi$
- `Sincerity on Intentions`:
  $\mathcal{P}_{a,D}\mathcal{I}_a\phi > \sharp_a\phi$

Together with axioms that link various speech act types to their illocutionary purpose and an axiom of `Cooperativity` ($\mathcal{P}_{a,D}\mathcal{I}_a\phi > \mathcal{I}_b\phi$; see below), these axioms ensure that the intentions behind a's current announcement become by default the intentions of all agents in D. Thus what one agent says can affect another agent's subsequent behaviour. For instance, the axioms predict from (1a) that A intends B to commit to C is stupid; B does this by announcing (1b). The axioms also predict from (1c) that A intends C to commit to C is not stupid, but A's intentions regarding B are more complex. A may not intend that B commit to (1c), and `Intent to Ground`, being defeasible, is compatible with this.

3.1 Desires

We have linked dialogue content to public commitment and the latter to belief and intention. But dialogue influences and is influenced by desires as well, and practical reasoning suggests that intentions are a byproduct of desires and beliefs. More precisely, rational agents intend those actions that maximise *expected utility*—utility reflecting one's desires or preferences, and *expectations* being based on *beliefs* about future outcomes. Preferences are thus distinct from but related to intentions.[1] We now address how an agent's preferences interact with other attitudes and dialogue content.

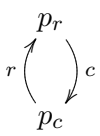
Games are a powerful model of preferences and actions among interacting agents. A game consists of a set of players and a set of strategies. Each strategy has a real-valued payoff or utility for each player. Typically the payoff for an individual is a function of *each* players' strategy, and intuitively, the payoff reflects that individual's preferences. A *Nash Equilibrium* (NE) is a combination of strategies that is optimal in that no player has a reason to deviate unilaterally from it. Games thus provide a method for computing one's next move in the dialogue. We illustrate this with a simple dialogue game in Table 4—a much simpler game than the ones that would underly the production of dialogues (1) to (3). In Table 4, R(ow) and C(olumn) are considering putdown moves (P_R and P_C) vs. praising moves. The cells indicate the utilities for agents R and C respectively for each combination of moves (e.g., column 2 row 2 defines the utilities for R and C when R praises C and C praises R). Note how the utilities for R and for C are influenced by what *both* agents do.

Since all utilities are defined, the game describes

[1] Preferences also have different logical properties: they can persist even after being realised while intentions don't; and they can be contrary to fact (one can prefer to be skiing right now while actually being at a meeting).

2/1	P_C	$\neg P_C$
P_R	$0, 0$	$3, -3$
$\neg P_R$	$-3, 3$	$4, 4$

Table 4: Simple Coordination Game



Preferences for C:
$\neg p_r \succ_c p_r$
p_r: $p_c \succ_c \neg p_c$
$\neg p_r$: $\neg p_c \succ_c p_c$

Preferences for R:
$\neg p_c \succ_r p_c$
p_c: $p_r \succ_r \neg p_r$
$\neg p_c$: $\neg p_r \succ_r p_r$

Figure 2: The CP-net for Table 4's 'Put Down' Game.

the complete preferences of each play with respect to all strategies. The two NEs are $(\neg P_R, \neg P_C)$ and (P_R, P_C). Utilities must be real values—standard game theory provides calculations of expected utility that combine probabilities over actions with the preferences for each player. But this sort of calculation is far too complex to be part of CL, which is a shallow logic for rough and ready decisions about discourse moves. To maintain a computationally effective CL, we need a *simpler* model of strategic reasoning that nevertheless *approximates* the types of interactions between expected moves and utility that game theory addresses.

Computationally efficient representations for strategic reasoning already exist. ***CP-nets*** (Boutilier et al., 2004) provide one such (qualitative) model for Boolean games (Bonzon, 2007)—games where like Table 4 each player controls propositional variables which he or she can make true or false (think of these as descriptions of actions that the agent performs, or not). A CP-net is designed to exploit the independence among the various conditions that affect an agent's preferences. It has two components: a directed *conditional preference graph* (CPG), which defines for each feature F its set of parent features $P(F)$ that affect the agent's preferences among the various values of F; and a *conditional preference table* (CPT), which specifies the agent's preferences over F's values for every combination of parent values from $P(F)$.

For example, the CP-net for the 'put down' game from Table 4 is shown in Figure 2. p_c stands for C doing a put down move; similarly for p_r. The dependencies among features for each agent are shown with labelled arcs in the CPG. The CPT then distinguishes among the conditional preferences for agents R and C; e.g., $\neg p_r : \neg p_c \succ_c p_c$ stipulates that C prefers not to put down R rather than put him down, if R does not put down C. The semantics of CP-nets ensures that its conditional *ceteris paribus* preferences generate a total order $\succeq$ over all possible combinations of values of all features. Roughly put, the logic of CP-nets adheres to the following two (ranked) principles when generating this total order: first, one prefers values that violate as few conditional preferences as possible; and second, violating a (conditional) preference on a parent feature is worse than violating the preference on a daughter feature. So the total preference orderings for R and C for the CP-net in Figure 2 are as follows:

$(\neg p_r \wedge \neg p_c) \succ_c (\neg p_r \wedge p_c) \succ_c (p_r \wedge p_c) \succ_c (p_r \wedge \neg p_c)$
$(\neg p_r \wedge \neg p_c) \succ_r (p_r \wedge \neg p_c) \succ_r (p_r \wedge p_c) \succ_r (\neg p_r \wedge p_c)$

In line with the game in Table 4, these orderings yield two NEs: $(\neg p_r \wedge \neg p_c)$ and $(p_r \wedge p_c)$. While there are games whose CP-net representations do not validate *all* the game's NEs, Bonzon (2007) shows that CP-nets predict all NE when quite general conditions on the games are met.

Unfortunately, it is an inescapable fact that the preferences of other agents are hidden to us: one estimates them from their actions, including their utterances. CL must therefore use information from the dialogue to infer the CP-net for agents; CL must also make use of partial or underspecified CP-nets. For instance, what R knows about C and *vice versa* will determine how they should 'play' the 'Put down' game. If R has the preferences from Figure 2, but C is a jerk—in other words, his preference is to play a putdown move, *whatever* the circumstances (so in contrast to Figure 2, his CPG contains no dependencies on p_c and his CPT is simply $p_c \succ_c \neg p_c$)—then this revised CP-net has a different NE; namely, $p_r \wedge p_c$. So, using the general strategy that R should choose a future dialogue move according to NE, he will do p_r. If, on the other hand C is not a jerk, with the CP-net from Figure 2, then R should play $\neg p_r$. So if R doesn't know if C is a jerk or a non-jerk, he can't guarantee his next move to be optimal. Such put-down games might therefore be useful for establishing what sort of person one is dealing with. R might engage in this game to

see how C acts (is C a jerk, or a non-jerk?), before R makes conversational moves towards other ends where the penalties are much higher.

3.2 Back to Cognitive Logic

As shown in Lang et al. (2003), one can translate CP-nets into a conditional logic. We can do the same with the weak conditional $>$ from CL. Our representation of a conditional preference in terms of $>$ introduces a predicate OK that labels a world as being a good outcome (Asher and Bonevac, 2005), where OK is always strictly preferred to $\neg OK$. We then adopt the following definition of agent a's conditional preference $\phi : \psi \succ_a \neg\psi$:

- `Preference in CL:` $(\phi : \psi \succ_a \neg\psi) \Leftrightarrow$
  $\phi \rightarrow (\neg((\phi \wedge \psi) > \neg OK_a) \wedge$
  $((\phi \wedge \neg\psi) > \neg OK_a))$

In words, some normal $\phi \wedge \psi$ worlds are better than all normal $\phi \wedge \neg\psi$ worlds. The unconditional preference $\psi \succ_a \neg\psi$ is thus $\neg(\psi > \neg OK_a) \wedge (\neg\psi > \neg OK_a)$. In contrast to reasoning with games and CP-nets directly, `Preference in CL` allows CL to reason with *partial* information about the relative preferences among all possible actions.

Let's now investigate how preferences link to other attitudes. First, there is a rationality constraint linking preferences to intentions. Consider an unconditional preference first:

- `Preferences to Intentions:`
  $(\phi \succ_a \neg\phi \wedge \mathcal{B}_a \Diamond_G \phi) > \sharp_a \phi$

In words, if an agent, all things considered, prefers ϕ and believes there to be a strategy for achieving ϕ in the contextually supplied game or decision problem G (our gloss for $\Diamond_G$), then defeasibly he forms the intention to ϕ. Preferences within a game allow us with `Preferences to Intentions` to specify a version of what Asher and Lascarides (2003) call the *Practical Syllogism* (PS), which links beliefs, intentions and the *choice* that marks one's preferred way of achieving goals.[2] Suppose G has a unique optimal solution s for agent a such that $s > \phi$. Then a prefers the sequence of moves leading to s to any alternative sequence, and by `Preferences to Intentions` that sequence is intended. Asher and Lascarides (2003) used PS to infer an agent's beliefs and intentions from his behaviour and *vice versa*. We can now do this without PS as a separate principle.

On the other hand, when speakers *publicly commit* to a certain intention or to a preference, then this is an at least defeasible sign about their *actual* preferences. So when reasoning about an agent, if he commits to a certain intention or a certain preference, this licenses a dynamic update of one's model of his preferences ($\heartsuit$ is the 'preferences' action operator, where $\heartsuit_a \chi$ effects a model transition where conditional preference χ is added to a's preferences, so long as it is consistent to do so):

- `Commitments to Preferences:`
  $(\mathcal{P}_{a,D}\mathcal{I}_a\phi \vee [\mathcal{P}_{a,D}(\phi \succ_a \neg\phi)]) >$
  $\heartsuit_a(\phi \succ_a \neg\phi)$

In cooperative games, it seems reasonable to suppose that in general if one agent prefers a certain outcome then so does another. That is, $(\phi \succ_a \psi) > (\phi \succ_b \psi)$ for players a, b in a cooperative game. This allows us together with `Preferences to Intentions` and `Commitments to Preferences` to derive the follow Cooperativity axiom:

- `Cooperativity:` $\mathcal{P}_{a,D}\mathcal{I}_a\phi > \mathcal{I}_b\phi$

Thus by using CP-nets and their translation into CL, we can deepen the foundations of CL itself, rendering more transparent the axioms assumed there.

We can also now make dynamic the interaction between information about cognitive states and dialogue moves. For example, let's examine R and C playing the putdown game in three scenarios that vary on how partial (or complete) R's and C's knowledge of each other's preferences are. First, suppose R and C have complete (and accurate) knowledge of each others preferences, which are those in Figure 2. Then by `Preferences to Intentions` R will intend $\neg p_r$ (i.e., praise C), and similarly C will intend $\neg p_c$ (i.e., praise R). By `Intent to Ground` both intentions will become also mutual intentions of R and C. And both have a rational expectation for how the verbal exchange

[2] They state PS as follows:
$(\mathcal{I}_a(\psi) \wedge \mathcal{B}_a((\phi > \psi) \wedge \mathit{choice}_a(\phi, \psi)) > \mathcal{I}_a(\phi)$
In words, if a intends that ψ, and he believes that ϕ normally leads to ψ and moreover ϕ is a's choice for achieving ψ, then normally a intends that ϕ. By treating the relation choice_a as primitive, the CL lacked the reasoning that agents engage in for finding optimal ways of achieving goals. We remedy this here.

will go.

Now consider the case where R's preferences are those in Figure 2 but R does not know if C is a jerk or not. On the other hand, C believes his own and R's preferences to be those given in Figure 2. Then R may not yet have formed an intention with respect to the goal, since he has no information on C's preferences or intentions. But C will act as above and thus R will learn about C's actual intentions. That is, on observing C perform $\neg p_c$ R will know that C intended it,[3] and by `Commitments to Preferences` she will update her model of C's preferences with $\neg p_c \succ_c p_c$. This now allows her to use the CP-net so-constructed to make the move that maximises her preferences—i.e., $\neg p_r$.

Finally, consider the case where R and C meet for the first time and don't know anything about each other's preferences. If R is to make the first move, then unlike the prior case R cannot use C's actions to influence her move. Instead, she must reason by 'cases', using each CP-net that is compatible with her own preferences. Suppose that R's preferences are those in Figure 2, and furthermore, R knows C to be either a non-jerk (as in Figure 2) or a jerk (making C's CP-net simply $p_c \succ_c \neg p_c$). Then R can reason as follows. If C is a non-jerk, then C prefers $\neg p_c$ on condition that R performs a $\neg p_r$ (reasoning as before), making R's best move $\neg p_r$. On the other hand, if C is a jerk, then C prefers p_c regardless, making R's best move p_r. R would therefore require further strategies for deciding which of p_r vs. $\neg p_r$ to prefer. For instance, R might 'hope for the best' and perform $\neg p_r$. In any case, where all that is involved is an insult, R may consider it better to potentially receive an insult and know about C's desires than to behave like a jerk herself. An extension of the CP-net could model these additional preferences.

4 Conclusions

In this paper we developed a cognitive logic for discourse interpretation that extends dynamic logics of public announcement. The extensions provide default links between public announcements and cognitive attitudes. It validates that grounding normally leads to mutual belief, but not always (see (1)). We also argued for representing preferences as $>$-statements, and highlighted the relationship between this and CP-nets—a compact way of representing Boolean games of the kind that have been used to model dialogue strategies. We thus linked within CL game-theoretic principles to general axioms of rationality and cooperativity. This affords a 'generate-and-test' way of deciding one's next dialogue move, even when one has only partial information about another agent's preferences. In future work, we plan to explore how to use this CL to model calculable implicatures (Grice, 1975).

[3] See Asher and Lascarides (2003) for details.

References

L. Amgoud. A formal framework for handling conflicting desires. In *Proceedings of ECSQARU*, 2003.

N. Asher and D. Bonevac. Free choice permission is strong permission. *Synthese*, pages 22–43, 2005.

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

A. Baltag, L.S. Moss, and S. Solecki. The logic of public announcements, common knowledge and private suspicions. Technical Report SEN-R9922, Centrum voor Wiskunde en Informatica, 1999.

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

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

B. Gaudou, A. Herzig, and D. Longin. Grounding and the expression of belief. 2006.

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

B. Grosz and C. Sidner. Plans for discourse. In J. Morgan P. R. Cohen and M. Pollack, editors, *Intentions in Communication*, pages 365–388. MIT Press, 1990.

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

J. Lang, L. van der Torre, and E. Weydert. Hidden uncertainty in the logical representation of desires. In *Proceedings IJCAI*, pages 685–690, 2003.

A. Lascarides and N. Asher. Agreements and disputes in dialogue. Proceedings of SIGDIAL, 2008.

D. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994.