Credibility and its Attacks

Antoine Venant[1], **Nicholas Asher**[2] **and Cedric Dégremont**[1*]
[1]IRIT, Université Paul Sabatier and [2]CNRS, IRIT

1 Introduction

Dialogues occurring in non-cooperative settings often exhibit attempts at deception, such as misdirections or lies. In many contexts, such as, e.g., trials or political debates, the objectives of a conversation's participant cannot be expressed in terms of her and her opponent's beliefs toward the content of the different dialog moves. In such contexts, dialogues moves come with semantic commitments of their own and challenges based on other players' semantic commitments.

In a political debate, an agent A might ask a question to another agent B, even though A knows the answer to the question. In such a case A is just seeking for B's commitment to an answer. If B complies and provides an answer, it can be in A's interest to further challenge this answer, even knowing it is correct. What is crucial here, are the objective semantic commitments that agents can force out of each other, rather than the subjective beliefs of these agents about whether the content of these commitments actually occurs or not.

Addressing the above requires us: 1. to have a semantic theory of commitments in dialogues, 2. to determine semantically what constitutes an attack and 3. to distinguish between attacks from a semantic perspective.

In the next section, we define credibility more precisely and attacks on it, linking these to linguistic commitments. In section 3, we give some examples of attacks on credibility, while sections 4 and 5 flesh out the analysis. Section 6 describes related work. We conclude with some directions for the future in section 7.

2 Credibility and commitments

An attack on credibility can be thought of as exposing deceitful intention. But determining intentions behind speech acts is a tricky business [14] we will not be getting into. The notions of credibility and attacks we are considering depends on overt and public linguistic commitments by speakers.

Using commitments, we now precise our notion of credibility: a dialogue agent i is not credible iff (i) it is shown for some φ that i has committed to φ that is absurd or clearly refutable (shown to be inconsistent with a prior claim of the agent or a background common assumption), and that it was plausibly in i's interest to commit to φ if φ is not attacked. An *attack* by player j on the credibility of i occurs iff j commits to the following: i has committed to φ, $\varphi \models \psi$, ψ is absurd or refutable, and it is in o's interest to commit to ψ, if ψ is not attacked. A move a by player i *makes possible* an attack on credibility iff it is discourse coherent for j to attach an attack on i's credibility to a.

Our notion of credibility differs considerably from that employed in the signaling games literature where credibility is defined in terms of beliefs, typically in equilibrium [10, 11]. Our notion of credibility is defined in terms of commitments, agent's interests and logical consequence, none of which depend on how the message affects the agents' beliefs.

To flesh out our picture of credibility and attack, we need to explain our notions of consequence and interest or preference. We have two notions of consequence: ordinary, logical consequence and defeasible consequence. We will assume that our agents are logically (though not factually) omniscient and so if i commits to φ he publicly also commits to ψ if ψ is a logical consequence of φ (notation $\varphi \models \psi$). Agents also commit to implicatures that are defeasible but what we shall term normal consequences that interlocutors would draw upon learning that i commits to φ. Finally, implicatures may be more tentative, as when i draws attention to an alternative to some-

[*] We thank ERC Research Grant 269427 for research support

thing to which he is explicitly committed. We'll assume that implicatures are modeled in a defeasible logic using a space of preferred models of the conversation. We also allow that some weak implicatures may exist only in some of the preferred models while stronger ones are true in all preferred models. We thus distinguish between the following three levels of commitment. -**Non-defeasible commitment by i to φ**: φ is a logical consequence of every possible interpretation of i's contribution. -**Implicit defeasible commitment by i to φ**: the "prefered" interpretations of i's contribution entail φ. -**Weak implicit defeasible commitment by i to φ:** some interpretations of i's contribution imply φ. Section 4 will provide more formal definitions.

We take preferences to be tied to a conception of rationality. In our framework, we will assume two conversational partners $0, 1$ and a third party who observes and judges but does not participate in linguistic exchanges. We will refer to this third party as *the jury*. The jury should be thought off as an abstract procedural entity with an objective look on the conversation that serves in the process of modeling rationality. The jury's mecanisms also depend on some contextual parameters which are common knowledge among the players: for instance, at court, the jury should know that an "honnest" expert witness must not share interest with the defendent lawyer. To give another example, he should know also when some facts are irrefutable and known as such. With this notion of a jury, our players prefer moves which make them look good in the eyes of the jury and make the other look bad, or at least worse. An attack by i on a player j's credibility is a way to make j look less good. Part of i's looking good is to not make mistakes, to not invite attacks on her credibility, but to make herself look good a player must provide positive reasons for the position she favors. *Mutatis mutandis* for the preferences of player j. More generally: (i) our players must play moves that make them look good; (ii) if player i is rational, she will prefer moves that make possible moves that j cannot attack; (iii) between 2 moves that make i look good but make possible attacks, she will prefer the one with the more indirect or weaker damaging context, since a more indirect damaging consequence is one that has a rebuttal move *that's not what I meant to say*.

3 Linguistic examples and intuitions

In this section we offer some linguistic examples featuring different sorts of commitments and attacks on credibility. These examples involve not only commitments to propositions expressed by assertoric clauses but also to propositions involving *rhetorical relations* that link clauses, sentences and larger units together into a coherent whole. That is, players commit to a particular content *and* to its relations with what has been said before. In so doing a player may also commit to contents proferred by his conversational partner as in [1].

Consider a case in which speaker A takes C's initial moves to be ambiguous.

(1) a. C: N. isn't coming to the meeting. It's been cancelled.
b. A: Did you mean that N. isn't coming because the meeting's cancelled or that the meeting is cancelled as a result?
c. C: As a result.

A's clarification question in (1)b presupposes that C's initial contribution was ambiguous between a result and an explanation move [7, 16]. We take this to imply at least a weak implicature for both readings, either of which a conversational participant could have exploited. This is something we want to model, and we'll see in the next example how such implicatures are exploited by an interlocutor.

Now consider the following example:

(2) a. C: N. isn't coming to the meeting. It's been cancelled.
b. A: That's not why N. isn't coming. He's sick.
c. C: I didn't say that N. wasn't coming because the meeting was cancelled. The meeting is cancelled because N. isn't coming.

This example illustrates how commitments embed. In (2)b A commits to the fact that C committed in (2)a to providing an explanation for why N isn't coming, even though (2)a is ambiguous. Only such a commitment explains why A attacks that commitment in the way that he does by giving an alternative explanation. But in fact, C takes that commitment by A to have misinterpreted him; C commits in (2)c that he committed in (2)a to offering a consequence or result of N's not coming to the meeting.

Note that while A attacks a move of C's in (2), he does not attack C's credibility in our sense. But neither does (2) provide a case of misleading implicature. However, the following example from [3] does. During the Dan Quayle-Lloyd Bentsen Vice-Presidential debate of 1988, Quayle was repeatedly questioned about his experience and his qualifications to be President. Quayle's attempted to compare his experience to the young John Kennedy's (referred to below as *Jack Kennedy* to convey familiarity) in his answer.

(3) a. Quayle: ... the question you're asking is, "What kind of qualifications does Dan Quayle have to be president," [...] I have as much experience in the Congress as Jack Kennedy did when he sought the presidency.
 b. Bensten: Senator, I served with Jack Kennedy. I knew Jack Kennedy. Jack Kennedy was a friend of mine. Senator, you're no Jack Kennedy.

Implicatures play a key role in his example. Quayle argues, against the thesis that his little governmental experience would make him unsuitable for the presidency, that Kennedy before him, with as much experience as he have, was able to handle the presidency. But this answer to the question suggests an implicit comparison between the two politicians (both junior senators from a state, each with little governmental experience) and gives rise to the possibility of interpreting Quayle's move as a stronger commitment that he would likely be able to handle the presidency in the same way that John Kennedy handled his, which, if not challenged would serve Quayle's claim better. Bentsen seized upon this weak implicature of Quayle's contribution and refuted it, indirectly exposing to the audience the self-serving nature of the comparison.

Here's an attested example from [17], in which a prosecutor (P) wants Bronston (B) to say whether he had a bank account in Switzerland or not, and Bronston does not want to make such a commitment for strategic reasons. But he defeasibly commits to an answer with (4)d in an attempt to avoid further questioning [2].

(4) a. P: Do you have any bank accounts in Swiss banks, Mr. Bronston?
 b. B: No, sir.
 c. P: Have you ever?
 d. B: The company had an account there for about six months, in Zurich.

It is interesting to consider a continuation of this in which the prosecutor would indirectly attack this response in (4)d.

(5) Prosecutor: I would like to know whether you personally ever had an account there?

If Bronston is forced on the threat of perjury to answer affirmatively, his response in (4)d now looks pretty deceiving to the Jury. The natural thought arises: Bronston was trying to deceive us into thinking that he didn't have an account. Though the prosecutor didn't proceed as in (5), had he done so he would have successfully attacked Bronston's credibility.

For our final example, consider the following excerpt from a *voir dire* examination in [12]. As background, the plaintiff lawyer (LP) has been repeatedly coming back to questions about the division of a nerve during a surgery with the objective of getting the witness (D) to characterize the surgical operation as incompetent and mishandled. Repeatedly coming back to the topic wore D down, and the defense attorney (LD) was no help:

(6) a. LP: And we know in addition to that, that Dr. Tzeng tore apart this medial antebrachial cutaneous nerve?
 b. D: Correct.
 c. LD: Objection.
 d. THE COURT: Overruled.
 e. D: Correct. There was a division of that nerve. I'm not sure I would say tore apart would be the word that I would use.
 f. LP: Oh, there you go. You're getting a hint from your lawyer over here, so do you want to retract what you're saying?

The defendant was resisting LP's line of attack relatively well, but then made an error by agreeing to LP's loaded question, in which LP makes the proposition that is really at issue, that Dr. Tzeng was negligent, a presupposition by embedding it under a factive verb. This makes it difficult to answer for D the question in a straightforward way.

Since D had already repeatedly been asked about this issue, he wasn't paying attention. LP successfully attacks D's credibility in (6)f when D attempts to correct his mistake with (6)e, by seizing on a weakly implicated discourse connection between (6)c and (6)e of Result* (the commitment in (6)c caused the commitment in (6)e).

These examples suggest two general methods of deception: moves that implicate propositions that can't be committed to explicitly for strategic reasons, and moves that trap agents into making commitments they should rationally refrain from.

Another feature of attacks is that generally they work gradually in damaging an opponent's credibility. Perhaps no one move succeeds on its own in convincing the jury that the opponent is duplicitous or incompetent; rather a series of moves gradually move a jury to a skeptical view of the opponent over the course of a conversation. The victory conditions for our players are to succeed in eventually moving the jury to a position in which the opponent is no longer credible.

4 Dialogue model

We need a dialogue model in order to analyze our examples and attacks on credibility in more detail. We've already seen that we need to model as part of a speaker's contribution not only its compositional semantics but also its illocutionary effects, in particular the implicit discourse links between utterances, as these can trigger or convey attacks on credibility. We will therefore build on [15], as SDRT already offers a formal, logic-based approach of dialogue content (semantics + illocutionary effects).

[15] models the semantics of dialogue by assigning to each conversational agents a *commitment slate*. Each commitment slate contains a list of propositions that an agent is committed to, which involve rhetorical relations as well as elementary propositions. [15] model explicit and implicit agreements and denials of one agent about another agent's commitments. However, the analysis of credibility threats requires that we go a step further. Conversational agents explicitly or implicitly refer to, and dispute, others' commitments. They attack their opponent's credibility by exposing inconsistencies in something they claim the opponent committed to or implicated, and defend against such attacks by denying a commitment to content that the opponent claims they committed to or implicated. We need to represent the commitments of all speakers from their own and their interlocutors' points of view, as in [18]. Moreover, we need to represent arbitrary nesting of commitments explicitly. Recall example (2). In (2)b A corrects C's prior utterance, and thereby commits that C is committed to a false proposition p (N. is not coming because the meeting is cancelled). C rejects A's correction. But what C rejects is not the proposition that corrects p, but A's commitment that C commited to p. Therefore, C also commits that A commits that C commits that p. Further, we need to distinguish between weak and strong commitments: when an agent tries to misdirect another, he might for instance give a weak commitment the look of a stronger one. Thus our dialogue model will add three things to [15]: explicit nested commitments, the commitments of each agent from every agents' point of view and explicit strong and weak commitments.

Conversations proceed as follows in our model: speakers alternate turns, each performing a sequence of discourse moves. Because we are interested in commitments and attacks, we will not import the full machinery of SDRT here. We will symbolize clausal contents within a propositional language, but incorporate labels for speech acts and discourse relations so that we can roughly express discourse-structures following [1]. Crucially, however, our language allows us to embed discourse structures under 3 modal operators $[\ ]$, $\langle\ \rangle$ and N. A discourse move for an agent i is defined as a discourse-level proposition labelled by a speech act identifier. A discourse-level proposition is either a base-level proposition, a formula expressing commitment over a discourse structure (*i.e.* i commits that a label have some particular content), or a complex formula $R(\pi_1, \pi_2)$ where R is a coherence-relation symbol and π_1 and π_2 are speech act labels. A complex formula recursively involves previously introduced speech acts labels. The modalities make the language more expressive, since we can express commitments of different agents to different contents for a single speech-act. The formula $[\pi : \gamma]_i$ states that agent i commits that the content of the speech act π is γ. Hence, she also commits that the speaker of π commits to the discourse proposition γ. Its dual, $\langle \pi : \gamma \rangle_i$, expresses the proposition that it is possible for i that the content of π is γ. $N_i \varphi$ means that i defeasibly commits to the contents of the formula φ. These modal operators express commitments over *discourse structures*. From this we retrieve commitments over *informational content* by looking at the content assigned to labels which are maximal for a given speaker. (labels that are not in the scope of another label of the same speaker): a speaker is committed to a content φ iff she commits the content of one of his maximal labels to be a proposition that entails φ.

Assume a set Φ of base-level propositions, a countably-infinite set of labels Π, a finite set of relation symbols $\mathcal{R}$ and a set of conversational agents I. In order to keep track of which agent x perform which speech act π, we assume Π partitioned in $|I|$ disjoint subsets $(\Pi_i)_{i \in I}$. We define spk(π) = the unique $i \in X$ such that $\pi \in \Pi_i$.

$$\Gamma(\Phi) := \varphi \mid R(\pi_1, \pi_2) \mid [\delta]_i \mid N_i(\delta) \mid \langle\delta\rangle_i \mid \neg\gamma$$

$$\Delta := \pi : \gamma \mid \pi :?(\gamma) \mid \delta_1 \wedge \delta_2$$

Where the γ_i and δ_i respectively range over $\Gamma(\Phi)$ and Δ, the π_i and φ_i respectively range over Φ and Π, and i and R respectively range over I and $\mathcal{R}$.

Definition 1 (Model). *A model $\mathcal{M}$ is a tuple $\langle W, v, (\rhd_x)_{x \in X}, < \rangle$, where W denotes set of possible worlds, $v : \Phi \mapsto \wp(W)$ a coloration, $<$: $W \to W^2$ a function from worlds to partial orderings over W, and for each agent x, $\rhd_x \subseteq W \times W$ is a transitive and euclidean accessibility relation.*

Our language has a dynamic semantics: the interpretation of a formula is context-change potential *i.e.* a relation between world-assigment pairs (w, σ). To account for polar question in our examples, we adopt a simplistic version of [13] and take propositions to semantically denote a set of set of worlds (a proposition denotes a set of possibilities which is partitioned into equivalence classes raised by questions). For intance, the question *whether p?* partitions a set of world in two, those worlds at which p on the one hand, and those at which $\neg p$ on the other. An assigment $\sigma : \Pi \times W \mapsto \wp(\wp(W))$ is a function that assigns a proposition as a set of set of worlds to a speech act label at a particular world. $\sigma(w, \pi)$ is roughly the (partitioned) set of worlds in which the interpretation of π at world w is true. Given a model $\mathcal{M}$, the function $[\![\cdot]\!]_{\mathcal{M}}$ maps each formula δ of the language to a binary relation $[\![\delta]\!]_{\mathcal{M}}$ over world-assignment pairs. Discourse-level assertoric propositions in $\Gamma(\varphi)$ always leave the assignment component unchanged and act as filters that let through only the worlds at which the proposition is true. Discourse moves in Δ on the other end modify the assignment. Another bit of needed machinery is for interpreting discourse relations. In our semantics each relation affects the contents assigned to its terms. Veridical relations like Explanation or Result will simply update the contextually given values to its terms with the semantic effects of the relation on those terms [1]. Non veridical relations like Correction or alternation place constraints on the truth of the contents associated with the terms at worlds verifying the relation in question. We need some notation first: assume a model $\mathcal{M} = \langle W, v, s, (\rhd_x)_{x \in X}, < \rangle$. Let p denote a dynamic proposition (*i.e* a relation between world/assigments pairs). Define $|p|^{\sigma}_{\mathcal{M}}$ as $\{w \in W \mid (\sigma, w)\ p\ (\sigma, w)\}$ and $|?p|^{\sigma}_{\mathcal{M}}$ as $\{|p|^{\sigma}_{\mathcal{M}}, W \setminus |p|^{\sigma}_{\mathcal{M}}\}$. Define $Acc(w)$ as the set of set of world containing a single element which is the set of all worlds accessible from w: $Acc_x(w) = \{\{w' \mid w \rhd_x w'\}\}$. Finally define the update operation $\star : \wp(\wp(W)) \times \wp(\wp(W)) \mapsto \wp(\wp(W))$ as $a \star b = \{x \cap y \colon x \in a \wedge y \in b\}$.

Definition 2 (Semantics). *Discourse propositions:*

$$(w, \sigma)[\![\varphi]\!]_{\mathcal{M}}(w', \sigma') \text{ iff } \begin{cases} (\sigma, w) = (\sigma', w') \\ w \in v(\varphi) \end{cases}$$

$$(w, \sigma)[\![R(\pi_1, \pi_2)]\!]_{\mathcal{M}}(w', \sigma') \text{ iff } (\sigma, w) = (\sigma', w') \text{ and } w \in I_R(\sigma(\pi_1, w), \sigma(\pi_2, w))$$

$$(\sigma, w)\ [\![[\delta]_x]\!]_{\mathcal{M}}\ (\sigma', w') \text{ iff } w = w' \text{ and } \forall w''\ w \rhd_x w'' \to (\sigma, w'')[\![\delta]\!]_{\mathcal{M}}(\sigma', w'')$$

$$(\sigma, w)\ [\![\langle\delta\rangle_x]\!]_{\mathcal{M}}\ (\sigma', w') \text{ iff } w = w' \text{ and } \exists w''\ w \rhd_x w'' \wedge (\sigma, w'')[\![\delta]\!]_{\mathcal{M}}(\sigma', w'')$$

$$(\sigma, w)\ [\![N_x \delta]\!]_{\mathcal{M}}\ (\sigma', w') \text{ iff } w = w' \text{ and } \forall u\, (w \rhd_x u \wedge \forall v (w \rhd_x v \to u \geq_w v)) \to (\sigma, u)\ [\![\delta]\!]_{\mathcal{M}}\ (\sigma', u)$$

Discourse moves:

$$(\sigma, w)\ [\![\pi : \gamma]\!]_{\mathcal{M}}\ (\sigma', w') \text{ iff } w = w' \text{ and } \sigma'(\pi, w) = \sigma(\pi, w) \star |\gamma|^{\sigma}_{\mathcal{M}} \star Acc_{\text{spk}(\pi)}(w)$$

$$(\sigma, w)\ [\![\delta_1 \wedge \delta_2]\!]_{\mathcal{M}}\ (\sigma', w') \text{ iff } w = w' \text{ and } (\sigma, w)\ [\![\delta_1]\!]_{\mathcal{M}} \circ [\![\delta_1]\!]_{\mathcal{M}}\ (\sigma', w')$$

Armed with this semantics for formulas, we can now define the commitments of each agent i at every initial prefix (sequence of turns) in the conversation. Because commitments will depend on discourse structure, we define commitments at *maximal* labels in the logical forms for the turns (those that are not within the scope of any other label). Given a logical form for n conversational turns (or the whole conversation), we can define the commitment of the players:

Definition 3. $(\sigma, w)\ [\![C_i\varphi]\!]_{\mathcal{M}}\ (\sigma, w)$ *iff*

$$\exists\pi(\text{spk}(\pi, i) \wedge\ maximal(\pi) \\ \wedge \forall u(w \rhd_x u \rightarrow \sigma(\pi, v) \subseteq |\varphi|))$$

We thus have a dynamic picture of how speakers' commitments evolve throughout a conversation.

Examples revisited. We start with example (2). [15] would analyse the two first turns as in table 1

turn	C's SDRS	A's SDRS
(2-a)	$\pi_1 : \neg N$ $\pi_2 : \mathit{ccl\_meeting}$ $\pi_3 : \mathit{Res}(\pi_1, \pi_2)$	
(2-b)		$\pi_4 : \neg \mathit{Exp}(\pi_1, \pi_2)$ $\pi_5 : \mathit{Corr}(\pi_3, \pi_4)$

Table 1: Analyisis of (2) following [15].

This is problematic, since A is committed to an absurdity. The semantic conditions of $Corr(\pi_3, \pi_4)$ require that the content of π_3 implies the negation of π_4, but $Res(\pi_1, \pi_2)$ does not imply $Exp(\pi_1, \pi_2)$ (the two are even contradictory). Keeping with the same kind of tabular representation as [15] our proposal amounts to further divide each cell of the table above in two, introducing A' interpretation of $C's$ moves, and repeating this process potentially infinitely to express arbitrary nestings as in table 2. For readability, we simplify the table by recopying at each step only the moves whose interpretation is controversial in the nested cells. In our language (2) is analysed as:

$$[\pi_1 : \neg N]_c \wedge [\pi_2 : \mathit{ccl\_meeting}]_c \\ \wedge [\pi_3 : Res(\pi_1, \pi_2)]_c \\ \wedge[\pi_4 : \neg Exp(\pi_1, \pi_3)]_a \wedge [\pi_5 : \langle\pi_3 : Exp(\pi_1, \pi_2)\rangle_c \\ \wedge\ \pi_5 : Corr(\pi_3, \pi_4)]_a \\ \wedge[[\pi_5 : \langle\pi_3 : Exp(\pi_1, \pi_2)\rangle_c]_a \\ \wedge\ \pi_6 : \neg C_x(Exp(\pi_1, \pi_2)) \wedge \pi_7 : Corr(\pi_5, \pi_6)]_c$$

Correcting move like π_5 triggers presupositions: here, a presupposition that c's move π_3 possibly commits him to the negation of π_4's content, accomodated as part of π_5's content. In the tabular representation, C's final move is:

C's SDRS		
C	A	
	C	A
π_6 : $\neg C_x(Ex(\pi_1, \pi_2))$ $\pi_7 : Cor(\pi_5, \pi_6)$	π_3 : $Exp(\pi_1, \pi_2)$	

In (2), we have only encountered explicit commitments $[\varphi]_x$. But in (1)b, A takes C's commitments to involve two possibilities, and he does not know which C has in fact committed to. Thus, in (1)b, A represents C's commitments as

$$[\pi_1 : \neg N]_c \wedge [\pi_2 : \mathit{ccl\_meeting}]_c \\ \wedge [\pi_3 : Res(\pi_1, \pi_2)]_c \\ \wedge[\pi_5 : \mathit{Clar\text{-}q}(\pi_4, \pi_3)]_a \wedge [\pi_5 : \langle\pi_3 : Exp(\pi_1, \pi_2)\rangle_c \\ \wedge\ \pi_5 : \langle\pi_3 : Res(\pi_1, \pi_2)\rangle_c]_a$$

In (3), Bentsen (B) seizes on a weak implicature of Quayle's (Q). Q explicitly commits to a direct comparison between his experience in government and that of the young JFK, but B corrects a more general equivalence between the presidential promise of JFK and his own. If we symbolize the latter with JFK, we take B's turn to yield $[\pi' : \neg\texttt{JFK} \wedge \langle\pi : \texttt{JFK}\rangle_q \wedge \pi_2 : Cor(\pi, \pi')]_b$. We see that even a weak implicature is sufficient to warrant B's corrective move in π_2. The success of this attack relies on the jury's decision on the admissibility of $\langle\pi : \texttt{JFK}\rangle_q$, *i.e.* the possibility of a commitent of q to JFK. Finally, in (6), we see that LP commits that D is committed to a discourse link between the defense attorney and his own self-correction:

$$[[\pi_1 :?p]_{lp} \wedge \pi_2 : p \quad \wedge\pi_3 : QAP(\pi_1, \pi_2)]_d \\ \wedge[\pi_5 : Obj(\pi_4, \pi_3) \wedge \pi_9 : \langle\pi_7 : Res(\pi_5, \pi_6)\rangle_d \\ \wedge\ \pi_9 : [\pi_8 : Cor(\pi_3, \pi_6)\rangle_d]_{lp}$$

C's SDRS	A's SDRS	
$\pi_1 : \neg N$ $\pi_2 : \mathit{ccl\_meeting}$ $\pi_3 : Res(\pi_1, \pi_2)$		
	C	A
	π_3 : $Exp(\pi_1, \pi_2)$	π_4 : $\neg Exp(\pi_1, \pi_2)$ π_5 : $Corr(\pi_3, \pi_4)$

Table 2: Adding nested commitments

5 The strategic model

Speakers choose the sequences of discourse moves they do because they want to convey commitments that will make them look good in the eyes of the Jury; they also want to make an opponent look bad if possible by attacking her weak points. We call this their winning condition. We will assume as in [3] that speakers may have incomplete knowledge of the other players' moves, leading to nasty surprises as in (3), where Quayle clearly didn't anticipate Bentsen's move in (3)b. A final desirable feature of the strategic model is that the moves open to a participant that lead her to her winning condition may decrease or even vanish if her credibility is repeatedly attacked. Thus the underlying framework of a sequential game is essential for analyzing conversation.

During play, a player has to weigh whether to make a move that makes her look good but that is risky in that it can be attacked; if the attack has no grounded rebuttal [9], the move could be disastrous. Further, when an opponent j makes a move involving an implicature it is up to the player i to decide whether it can be taken as a safe commitment in the sense of [2], and to exploit it in subsequent conversational moves, as the prosecutor of (4) does; and conversely j has to weigh whether the player will take the implicature on board or not, as one of i's commitments. If not, the deceptive move may fail if an opponent makes a request for an explicit version of an implicated commitment as in (5).

All of these calculations depend on the effect of play on the Jury, who ultimately decides the winner according to positive points and lack of bad moves (inconsistencies or deceptions) on the part of i and other players. Our Jury entertains a space of possibilities concerning player types for the players and a probability distribution P over them. Our model is simple; we assume just two types for each player GOOD and BAD. At the start of the conversation the Jury entertains only the possibility that all players are GOOD; that is the probability distribution is such that $P(\text{BAD}_i) = 1 - P(\text{GOOD}_i) = 0$ for any player i. As the conversation proceeds, $P(\text{BAD}_i)$ is successively updated given what has happened over the last turn; i.e. $P_n(\text{BAD}_i) = P_{n-1}(\text{BAD}_i/t_n)$. As long as the opponent does not convincingly refute the arguments of i at n, $P_n(\text{BAD}_i) = P_{n-1}(\text{BAD}_i)$. However, a successful attack on, say, i by j at turn t_n, which results in a refutation of an argument by i with no convincing rebuttal gets the Jury to update P via Bayesian conditionalisation such that $P_n(\text{BAD}_i) > P_{n-1}(\text{BAD}_i)$.

The effect of a higher probability on BAD_i is that the positive reasons advanced by i are given a lower score; that is the effect of a bad reputation—the good things you say get discounted. Thus, if the positive arguments by i in her favor provide some positive score σ^i, then the effect on the Jury at turn n is:[1]

$$\text{overall-score}_n^i = P_n(\text{GOOD}_i)(\sigma_n^i) \qquad (1)$$

Our model should also reflect the duplicitous nature of weak implicatures that agents don't dare put out as full commitments. So the update of the probability on $P_n(\text{BAD}_i)$ will depend on (i) the strength of the implicature, (ii) whether the attack is successful in so far as there is no rebuttal that refutes it. For weak implicatures, there is a rebuttal: you misinterpreted what I said; but it renders the move useless for the player. The upshot of our model is that agents pay dearly if their credibility is successfully attacked when they advance a weak implicature, as evidenced in the example (3). What exactly was Quayle's (DQ) mistake? It was that he weakly implicated that he was of the same caliber as JFK, and it is this implicature that Bentsen (B) seizes on and shows to be ridiculous. He also implicates that DQ insinuated the direct comparison without directly saying so, which is a deceptive move. B's attack was very successful, especially since DQ did not vigorously rejoin with *you misinterpreted me unfairly*. Our model considerably increases $P_n(\text{BAD}_i)$ conditional on B's attack, rendering DQ's successful points much weaker. Not only did B refute DQ's argument and expose his deceptive move, but he affected the overall outcome of the debate.

In example (4) on the other hand Bronston (B) carries off his strongly implicated commitment to an answer in (4)d without being challenged by the prosecutor (P). However, it was somewhat dangerous. Had P continued as in (5), B would have had to contradict $N_B(\neg\text{bank account})$ with $[\text{bank account}]_B$, the two modal formulas being inconsistent. Bronston could have claimed that he had not understood the first prosecutor's question as being directly about him, but the response

[1] In future work, we plan to experiment with refinements of this basic idea, such as using updated probabilities to score continuations.

would have been weak and our intuition is that his credibility would have suffered. Example (6) exhibits a slightly different pattern: D has committed to there being no negligence in the operation, but in (6)b, he commits to the presupposition of the question that entails negligence. Conditional on such a contradiction, $P(\text{BAD}_R)$ increases, but not much because it was a trick question. But when D attempts to retract his affirmative answer to LP's biased question, then pins on him reasoning that attacks his credibility as an impartial witness via a weakly implicated connection between LD's and D's contributions. At this point, $P(\text{BAD}_D)$ increases considerably, and weakens D's testimony in the eyes of the Jury.

Our model predicts that as a player's credibility is repeatedly attacked and duplicitous moves are exposed, her credibility decreases monotonically. As a consequence, after a certain point the player may have no moves open to her that achieve her winning condition—the probability of her being of BAD type is now too high.

6 Related work

Our work assumes a commitment based view of conversation rather than one based on the internal, mental states of the participants [14, 20, 19] and builds on and complements the model proposed in [18], which in turn extends [7]. They introduce a dynamic Bayesian model for discourse actions based on prior moves. Our paper is more limited in scope but also goes into more detail: our model details how attacks on credibility function with respect to various types of commitments that come from different kinds of discourse moves; we show that even in simple conversations levels of embedded commitments can be very complex (contrary to a suggestion of [18]); and our Bayesian update on player types details a part of the picture of [18].

Our model assumes a sequential game view of conversation, differing from extended signaling games [4], and uses a notion of credibility which differs from the standard one in signaling games, according to which a message is credible iff its standardly accepted content understood as a set of evaluation points is a superset of its meaning in reflective equilibrium (roughly how the message content affects belief). In strategic environments of the sort we have in mind, signaling games have severe limitations: in our strategic contexts, a player will send a message only if it benefits him, but then that message will not benefit his opponent. In a signaling game, the opponent should rationally ignore it [6]. However, in a debate, it would be irrational to ignore the message of the opponent. Our notion of credibility does not mix belief and action in the way signaling games do, and is immune to this problem. A further problem with signaling games is that they assume common knowledge of the preferences of each player over moves. But if these are used to define or to guide credibility, then there is no room for maneuver or deception, which is manifest in our examples. However, our model leaves a place for a signaling game analysis between the Jury and the players, which we will pursue in future work.

Related to our work are also recent attempts to investigate argumentation in actual dialogue [5]. Argumentation theory provides a framework for analyzing attacks and counterattacks [9]. We have given much more linguistic detail on how such attacks are carried out and how this can affect ones' strategy in conversation. On the other hand, we have presented a general model for credibility in strategic conversation. Different contexts may affect the parameters of the model that we have set up. For instance,[2] sometimes the Jury may be a participant in the conversation in the sense that it is allowed to ask questions, sometimes not. Given a particular context, Jury might also function according to persuasion rules that are different from the simple one we have used in section 5. We have chosen simple settings to illustrate our model. Finally, we have not gone into the details of how particular conversational contexts may dictate specific linguistic forms of attacks and defense, e.g., [8]. Our model is general enough, we believe, so that we can tune the parameters to fit the particularities of specific contexts.

7 Conclusions

In this paper, we have presented new notion of credibility and attacks on credibility that are relevant to conversations in strategic settings where interlocutor preferences may be opposed. We have developed a dialogue model extending both [15] and [18] with a semantics for dialogue turns and commitments that allows for arbitrary nestings of commitments. We have also shown that this complexity is required to analyze many examples of dialogue with attacks on credibility.

[2] Thanks to a Semdial reviewer for this point.

References

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

[2] N. Asher and A. Lascarides. Strategic conversation. *Semantics and Pragmatics*, Vol 6.2:1–62, 2013.

[3] N. Asher and S. Paul. Infinite games with uncertain moves. In F. Mogavero, A. Murano, and M. Vardi, editors, *Proceedings of the First Workshop on Strategic Reasoning, ETPCS 2013, Rome, Italy*, pages 25–32, Rome, Italy, 2013. Springer.

[4] R. Aumann and S. Hart. Long cheap talk. *Econometrica*, 71(6):1619–1660, 2003.

[5] Elena Cabrio, Sara Tonelli, and Serena Villata. A Natural Language Account for Argumentation Schemes. In *AI*IA - XIII Conference of the Italian Association for Artificial Intelligence - 2013*, Turin, Italie, December 2013. Springer.

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

[7] D. DeVault and M. Stone. Managing ambiguities across utterances in dialogue. In *Proceedings from the International Workshop on the Semantics and Pragmatics of Dialogue (DECALOG 2007)*, Trento, Italy, 2007.

[8] Paul Drew. Contested evidence in courtroom cross-examination: The case of a trial for rape. *Talk at work: Interaction in institutional settings*, pages 470–520, 1992.

[9] P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[10] Joseph Farrell. Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5(4):514–531, 1993.

[11] M. Franke, T. de Jager, and R. van Rooij. Relevance in cooperation and conflict. *Journal of Logic and Language*, 2009.

[12] R. Friedman and P. Malone. *Rules of the Road: A Plaintiff Lawyers Guide to Proving Liability*. Trial Guides, 2nd edition, 2010.

[13] J. Groenendijk and M. Stokhof. *Studies on the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, Centrale Interfaculteit, Amsterdam, 1984.

[14] C. Hamblin. *Imperatives*. Blackwells, 1987.

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

[16] M. Purver. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, Department of Computer Science, King's College, London, 2004.

[17] L.M. Solan and P.M. Tiersma. *Speaking of Crime: The Language of Criminal Justice*. University of Chicago Press, Chicago, IL, 2005.

[18] M. Stone and A. Lascarides. Grounding as implicature. In *Proceedings of the 14th SEMDIAL Workshop on the Semantics and Pragmatics of Dialogue*, pages 51–58, Poznan, 2010.

[19] D. Traum. Computational models of non-cooperative dialogue. In *Proceedings of the International Workshop on the Semantics and Pragmatics of Dialogue (LONDIAL)*, London, 2008.

[20] D. Traum and J. Allen. Discourse obligations in dialogue processing. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL94)*, pages 1–8, Las Cruces, New Mexico, 1994.