Strategic conversations under imperfect information: Epistemic Message Exchange Games

Nicholas Asher, Soumya Paul

Strategic conversations under imperfect information: Epistemic Message Exchange Games

Nicholas Asher and Soumya Paul

IRIT, Université Paul Sabatier
31062 Toulouse, France

Abstract. This paper refines the game theoretic analysis of conversations in [5] by adding epistemic concepts to make explicit the intuitive idea that conversationalists typically conceive of conversational strategies in a situation of imperfect information. This "epistemic" turn has important ramifications for linguistic analysis, and we illustrate our approach with a detailed treatment of linguistic examples.

1 Introduction

It has long been a common sense intuition of many philosophical and linguistic theories of communication that a conversational contribution should be interpreted in light of the participants' own beliefs and plans, in particular their beliefs about beliefs and plans of the other participants and so on. However, most formal theories of semantics and pragmatics do not rigorously implement this conception. Game theoretic analyses, in particular epistemic game theory and is notion of a type, provide the tools to do this and a new perspective on the notion of context and its role in the interpretation of ambiguous linguistic signals, as we show here. While some work has investigated the notion of meaning using games in cooperative settings [11], only the approach of [5] applies a game-theoretic approach to entire conversations. Moreover, this approach, which we build on here, does not require assumptions of cooperativity on which conversationalists have convergent interests. In this paper, we show how adding epistemic concepts to the approach in [5] yield compelling interpretations of dialogues.

Our paper also has a technical aim. [5] introduced a new framework for analyzing conversations in strategic settings, *Message Exchange (ME) games*. We fill here what we see as an important lacuna in that and subsequent work [4, 6, 7]. To date, although ME games have been analyzed in the setting of imperfect information, the information structure at work in the dynamics of the game has not been elucidated. We provide an analysis of ME games under assumptions of imperfect information, incorporating the beliefs of the players about their opponents and about a key component of ME games, the Jury, which is an abstract scoring device that assigns winning conditions to the players.

Our paper begins with a motivating example for adding epistemic concepts (Section 2), and then proceeds to provide some technical preliminaries in Section

3. We then go on to introduce the notions of types and belief functions and the tools required to develop them in our setting in Section 4. We have put more technical details fleshing out Sections 3 and 4 in the Appendix. We treat some linguistic examples in detail to show how the machinery works, and we then draw some conclusions for linguistic interpretation, its treatment of ambiguity, its subjectivity, and some sources for the biases inherent in that subjectivity. Section 5 offers some further thoughts on the subjectivity of information and conclude by mentioning some questions for future research that our new framework suggests.

2 Why add epistemic concepts to ME games?

ME games are infinitary games with two players, each playing a finite sequence of discourse moves, in alternating fashion. As the players make moves, the semantics of these moves entails certain public commitments on their part. The players need not have any interests or goals in common. Nevertheless, they conduct the conversation and pay attention to each others' public commitments in the hope of persuading the Jury to award them a win in the conversation.

As a motivating example, consider the following excerpt from a courtroom proceedings where a prosecutor is querying the defendant (originally from [16] and later discussed in [3, 5]).

Example 1 *Bronston and the Prosecutor:*

a. **Prosecutor***: Do you have any bank accounts in Swiss banks, Mr. Bronston?*
b. **Bronston***: No, sir.*
c. **Prosecutor***: Have you ever?*
d. **Bronston***: The company had an account there for about six months, in Zurich.*
e. **Prosecutor***: Thank you Mr. Bronston.*

For convenience, let us denote Prosecutor and Bronston with the letters P and B respectively. One conversational goal of the P in Example 1 is to get B to commit to an answer eventually (and admit to an incriminating fact) or to continue to refuse to answer (in which case he will be charged with contempt of court). While the conversation here is of course finite, we cannot say that P wins if B does not answer immediately. That would disallow questions of clarification, vague or imprecise responses that demand clarification and so on, as [5] argue. Nor is it plausible, given that we are interested in analyzing actual conversational practice, to fix in advance the number of such clarification questions or follow up questions and their responses in advance. Thus, the goals of P must be stated in such a way that they are naturally analyzed as conditions on *infinitary plays*, which are strings of discourse moves. Using the adverb *eventually* in our description shows that a natural way to think of P's winning condition is in terms of a formula of linear temporal logic or, equivalently, a first order formula over the language of infinite strings.

While we follow this analysis, which improves considerably upon [3], the account in [5] is still incomplete. In particular, the analysis of the epistemic

situation in Example 1 is only partial in [3] and non-existent in [5]. The epistemic situation of Bronston, the Prosecutor and the Jury are key to an analysis of this example. To understand the strategies of B and P in Example 1, it is natural to suppose that B has one belief about the Jury, P has another. For instance, why is P satisfied with B's answer? Presumably, in asking the question, P's goal was that B answer, or go on record as refusing to answer, the question. But B did not commit to having a bank account, or to not having a bank account, but he also did not refuse to answer the question and implicated a negative answer to the question. Why did P not follow up with a response like:

e'. *I'm not interested in what your company did, Mr. Bronston. I want to know whether you ever had a bank account at that institution.*

But now suppose that P believes of the Jury that having the implicated answer is sufficient; this counts as B committing to not having had a bank account. This would explain the lack of the continuation (e'.); P does not respond with (e'.), because he believes he has already attained his objective for this part of the conversation. In fact the Bronston trial concluded in P's favor in the lower court, and so his beliefs about the Jury were justified.

Now what about B? B could have simply been inattentive, but another epistemic assumption naturally explains his behavior of producing an indirect answer to P's question: B thought that the Jury might have a different standard for commitments than P did. If B assumed that the Jury is of the *type* according to which he could only be convicted on the basis of firm or hard, not implicated, commitments about his involvement with the Swiss bank, then an implicated response would be rational. If we suppose further that B believes that P might assign the Jury a type of the sort P actually seems to have, then B's response is optimal. First, the indirect response avoids making hard commitments that incriminate him according to his conception of the Jury, and secondly, the response satisfies P and so P does not probe further, in particular with potentially damaging continuations like (e'). In fact B's beliefs about the Jury, though they were not accurate for the lower court, were right concerning the Jury of the Supreme court, to which B and his lawyers appealed (and won) [16].

Example 1 will guide our analysis in this paper, although there are many other applications of epistemic concepts in ME games outside a courtroom setting where one, some of which are to be found in [5]. Epistemic concepts have a role to play in explaining the use of discourse moves and in terms of strategies for achieving an agent's conversational goals in any conversation where participants' interests are not completely aligned. Epistemic concepts are also crucial for technical reasons in ME games where players' interests are opposed but it is common knowledge that the conversation cannot really go on forever. We will formalize beliefs players have about other players and about the Jury using types in the manner of [12], as is standard in epistemic game theory [9, 14]. The types and beliefs that interest us here are beliefs about *linguistic moves* and propensities of agents to use certain linguistic moves and strategies. We then apply our formalism to standard ME games, and to discounted ME games [5, 4].

3 Preliminaries

In this section we develop the necessary background for the formal model to be presented in the following sections, re-visiting some of the definitions for ME games introduced in [5]. We then go on to introduce the notions of types and belief functions and the tools required to develop them in our setting. We will work in a setting where there are two players (conversationalists), 0 and 1, to simplify the notation and examples. The definitions and results easily generalize to settings with more than two players. In what follows i will always take a value in $\{0, 1\}$ unless mentioned otherwise. Player $(1 - i)$ will denote the opponent of Player i.

[5] defines ME game as infinite games over a countable 'vocabulary' V. The intuitive idea behind an ME game is that a conversation proceeds in turns where in each turn one of the players 'speaks' or plays a string of elements from V. In addition, in the case of conversations, it is essential to keep track of "who says what". To model this, each player i was assigned a copy V_i of the vocabulary V which is simply given as $V_i = V \times \{i\}$. Thus when Player i plays $u \in V$ (say), it is noted as (u, i). The formal vocabulary for our ME games is thus the set $(V_0 \cup V_1)$. The conversations will thus correspond to plays of ME games which are the union of finite or infinite sequences in $(V_0 \cup V_1)$, denoted as $(V_0 \cup V_1)^*$ and $(V_0 \cup V_1)^\omega$ respectively. The set of all possible conversations is thus $((V_0 \cup V_1)^* \cup (V_0 \cup V_1)^\omega)$ and is denoted as $(V_0 \cup V_1)^\infty$. As usual, we let $(V_0 \cup V_1)^+ = (V_0 \cup V_1)^* \setminus \{\epsilon\}$ where ϵ is the empty sequence.

Given any set A, we let $\Delta(A)$ be the set of all probability measures on A. Now, because our conversational agents will be have beliefs with a certain probability about what other agents may say in a given situation, we will need to endow the set of all possible conversations (plays), with a measure. In the Appendix, we show how to endow the set of finite and infinite strings over any generic non-empty set X with a topology. So, when we consider plays in our ME games, which will be sequences over $(V_0 \cup V_1)$, we shall assume that they have the topology defined for X.

3.1 The vocabulary

We now formally define the vocabulary V of an ME game. Players do not play just any sequence of arbitrary strings but sentences or sets of sentences that 'make sense'. To ensure this, the vocabulary V should have an exogenous semantics built-in. In order to achieve this, we exploit a semantic theory for discourse, SDRT [2]. SDRT develops a rich language to characterize the semantics and pragmatics of moves in dialogue. This means that we can exploit the notion of entailment associated with the language of SDRSs to track commitments of each player in an ME game. In particular, the language of SDRT features variables for dialogue moves that are characterized by contents that the move commits its speaker to. Crucially, some of this content involves predicates that denote rhetorical relations between moves—like the relation of *question answer pair* (qap), in

which one move answers a prior move characterized by a question. The vocabulary V of an ME game thus contains a countable distinguished set of individual constants or discourse constituent labels $\mathsf{DU} = \{\pi, \pi_1, \pi_2, \ldots\}$, and a finite set of discourse relation symbols $\mathbb{R} = \{\mathcal{R}, \mathcal{R}_1, \ldots \mathcal{R}_n\}$, and formulas $\phi, \phi_1, ...$ from some fixed language $\mathcal{L}$ for describing elementary discourse move contents, a language like that of higher order logic used in, e.g., in Montague Grammar. V consists of SDRT formulas of the form $\langle \pi : \phi \rangle$, where ϕ is either a formula of L, a relational formula of the form $\mathcal{R}(\pi_1, \pi_2)$, which says that π_1 stands in relation $\mathcal{R}$ to π_2 (one such relation $\mathcal{R}$ is qap), or a conjunction of SDRT formulas and relational formulas. When ϕ is a formula of L, then the DU π such that $\pi : \phi$ is called an *elementary discourse unit* or EDU; when ϕ is a conjunction of SDRT formulas and relational formulas, we say that $\pi : \phi$ is a *complex discourse unit* or CDU. Each discourse relation symbolized in V comes with constraints as to when it can be coherently used in context and when it cannot. Also note that since $\mathcal{L}$ is the language of some higher order logic, it can existentially quantify over the relation symbols in $\mathcal{R}$.

We can define an equivalence relation $\sim$ on V based on what coherent and consistent continuations they allow. $\phi_1 \sim \phi_2$, if for any SDRT formula ψ, $\phi_1.\psi$ is a consistent and coherent continuation just in case $\phi_2.\psi$ is, where coherence and consistency are defined as in [6]. If $\phi_1 \sim \phi_2$ then ϕ_1 and ϕ_2 are semantically equivalent, though the reverse is not true; ϕ_1 and ϕ_2 may have the same truth conditional content but give rise to different continuations in virtue, for example, of their politeness register. We shall refer to a $\sim$ equivalence class of V as a *class* of discourse moves. In the examples that follow, when we talk of a 'move', we shall actually be referring to its class.

3.2 ME game

We now define an ME game. In the definition we use a term $\mathcal{J}$ which stands for a 'Jury' of the ME game. It is the Jury who determines which player (or players) has achieved her goal in the conversation, or in other words, fixes the winning conditions for the players. We shall formally define the Jury in the following subsection.

Definition 1 (ME game [5]). *A* Message Exchange game (ME game)*, $\mathcal{G}$, is a tuple $((V_0 \cup V_1)^\infty, \mathcal{J})$ where $\mathcal{J}$ is a Jury.*

The ME game proceeds in turns where, by convention, Player 0 starts the game by playing x_1, Player 1 follows with x_2, Player 0 then plays x_3 and so on. This results in the sequence $x_1 x_2 x_3 \ldots$. Given our vocabulary V this sequence is a concatenation of formulas from $\mathcal{L}$ where concatenation is viewed as conjunction. As an example consider the following conversation between players 0 and 1.

Example 2 *Player 0 plays the sequence x_1 followed by Player 1 who plays the sequence x_2 where*

a. x_1 = *(Why did you come back?, 0)*

b. x_2 = *(The meeting has been cancelled. $\mathcal{N}$ did not turn up., 1)*

This results in the sequence x_1x_2 and so on. To cast this in terms of the language $\mathcal{L}$, let ϕ_1, ϕ_2 and ϕ_3 be the EDUs:

- ϕ_1 = *Why did you come back?*
- ϕ_2 = *The meeting has been cancelled.*
- ϕ_3 = *$\mathcal{N}$ did not turn up.*

ϕ_1 and ϕ_2 bear a clear semantic relation to each other—namely, that ϕ_2 is an answer to ϕ_1, written as $\mathsf{qap}(\pi_1, \pi_2)$ in the language $\mathcal{L}$. Such obvious semantic relations are part of the discourse grammar of the language and common knowledge of the participants. Work on discourse parsing as in [1] makes precise a notion of discourse grammar to which we will appeal here: the grammar is something that is learned and assigns a probability distribution to connections between discourse units and the relations that label those connections. On the other hand, the relation between π_3 and the previous EDUs is less clear; the grammar tells us that π_3 is connected to π_2 with high probability, but the grammar does not assign a high probability to just one relation holding between them. We could interpret the relation as π_3's furnishing an explanation of the event described in π_2—something we denote in $\mathcal{L}$ by $\mathsf{expl}(\pi_2, \pi_3)$. Or we could interpret the relation as π_3's providing a result or causal effect of the event described in π_2 (denoted in $\mathcal{L}$ by $\mathsf{res}(\pi_2, \pi_3)$). But because these relations have incompatible semantic consequences, both cannot apply to π_2 and π_3. In such a case, we say that the relation between π_2 and π_3 is underspecified *and we simply quantify over a relational variable to signify this. In the language of $\mathcal{L}$ the sequence x_1x_2 thus is:*

$$x_1x_2 \equiv (\langle \pi_1 : \phi_1 \rangle, 0)(\langle \langle \pi_2 : \phi_2 \rangle \mathsf{qap}(\pi_1\pi_2) \langle \pi_3 : \phi_3 \rangle \exists \mathcal{R} \cdot \mathcal{R}(\pi_2, \pi_3) \rangle, 1)$$

x_1x_2 is what [2] calls an *underspecified logical form* or ULF. The relation between π_2 and π_3 is underspecified, because the grammar (syntax, compositional and lexical semantics) does not determine it.

This leads us to the definition of a play of an ME game.

Definition 2 (Play). *A* play *ρ of an ME game is a sequence in $(V_0 \cup V_1)$.*

ρ can be a ULF. Once the existentially quantified relation variables are instantiated with specific relation names, we get what [2] calls a *fully specified logical form* or FLF. Thus, a single ULF can give rise to a (finite) number of FLFs or SDRSs, as, in the language of SDRT, an FLF is nothing but a complete SDRS. We define each such SDRS to be a history.

Definition 3 (History). *A* history h *of an ME game is an* FLF *or equivalently an SDRS.*

Given a play ρ, $\mathsf{h}(\rho)$ denotes the set of all histories generated by instantiating the existentially quantified relation variables in ρ with witnesses from the set of actual relation terms. Let $|\rho|$ denote the number of turns in a play ρ and $|\mathsf{h}|$

denote the same for the history h. By definition, $|\rho| = |\mathsf{h}|$ for all $\mathsf{h} \in \mathsf{h}(\rho)$. We let $\mathcal{P}$ (resp. $\mathcal{H}$) denote the set of all plays (resp. histories), where $\epsilon \in \mathcal{P}$ (resp. $\epsilon \in \mathcal{H}$) is the empty play (resp. empty history). We say that a play (resp. history) of the form $(V_0 \cup V_1)^+ V_i^+$ is an i-play (resp. i-history). We denote the set of i-plays (resp. i-histories) by $\mathcal{P}_i$ (resp. $\mathcal{H}_i$). Thus $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$ and $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$. Moreover, given a play ρ (resp. a history h) we let $\rho_j,\ 0 \leq j \leq |\rho|$ (resp. h_j) denote the length-j prefix of ρ (resp. h). That is, it is the play/history after j turns. Finally, given a play ρ of an ME game it will be convenient to define the $\mathsf{constituent}$ of ρ, $\mathsf{cons}(\rho)$, which is the sequence of DUs that occur in ρ in the same order.

Example 3 *Continuing with our Example 2, the play after two turns is* $\rho = x_1 x_2$. *The existential relation* $\mathcal{R}$ *in* ρ *can be interpreted in* at least *three different ways. This results in the following three histories (SDRSs/*FLF*s) [we supress the turn indices for better readability].*

a. $\langle \pi_1 : \phi_1 \rangle \langle \pi_2 : \phi_2 \rangle \langle \pi_3 : \phi_3 \rangle \mathsf{expl}(\pi_2, \pi_3)$
b. $\langle \pi_1 : \phi_1 \rangle \langle \pi_2 : \phi_2 \rangle \langle \pi_3 : \phi_3 \rangle \mathsf{res}(\pi_2, \pi_3)$
c. $\langle \pi_1 : \phi_1 \rangle \langle \pi_2 : \phi_2 \rangle \langle \pi_3 : \phi_3 \rangle$

Where the third interpretation is where the $\mathcal{R}$ *is interpreted as the vacuous relation. We shall represent a play which is an* ULF *as shaded regions in the ME game tree as can be seen in Figure 1 which is the game tree for the example 2.*

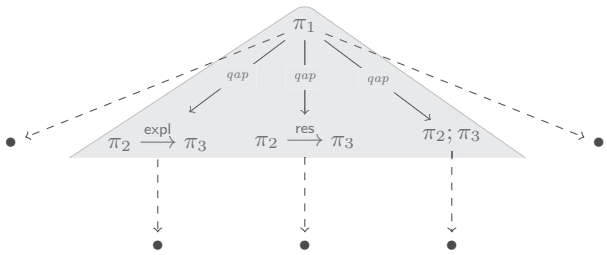

Fig. 1: The ME game tree for the conversation in Example 2

3.3 The Jury

We now formally define the concept of a Jury. The Jury of an ME game is an entity that 'evaluates' an instance of the game (a play which is the actual strategic conversation) and decides the winner after a 'finite number of turns' based on whether or not a player has, according to it, achieved her goal. For a finite sequence $x \in (V_0 \cup V_1)^*$, let $\mathcal{O}(x)$ denote the set of all (finite and infinite) continuations of x. That is, $\mathcal{O}(x) = \{xy \mid y \in (V_0 \cup V_1)^\infty\}$. In topological terms

$\mathcal{O}(x)$ is the 'basic open set' corresponding to x [for more on the topological aspects of sequences of conversations, see [5]]. We can then define

Definition 4 (Jury). *The* Jury *of an ME game is a tuple* $\mathcal{J} = (Win_0, Win_1)$ *where* $Win_i \subset (V_0 \cup V_1)^\infty$ *for each* i *and satisfies the following conditions:*

- **Finite checkability:** *for every finite sequence* $x \in (V_0 \cup V_1)^*$, $x \in Win_i$ *if and only if* $\mathcal{O}(x) \subset Win_i$.
- **Consistency:** *for every play* ρ *of the ME game,* $\rho \in Win_i$ *iff* $\mathsf{h}(\rho) \subset Win_i$.

Win_i *is called the* Jury winning condition *or simply the* winning condition *for Player* i.

The idea behind the restriction of finite checkability is that if after a certain finite sequence of exchange x, $x \in Win_i$ for some i then the Jury can already declare i as the winner of the conversation because it is assured that every continuation of x will be winning for i. If that is not the case, the Jury is still uncertain about the winner and has to let the conversation evolve further.

Definition 5 (Winning plays/histories). *A play* ρ *(resp. a history* h*) is said to be* winning *for Player* i *if* $\rho \in Win_i$ *(resp.* $\mathsf{h} \in Win_i$*).*[1]

Note that it might be that $(Win_0 \cap Win_1) \neq \emptyset$, in which case $x \in (Win_0 \cap Win_1)$ is winning for both players.

[5] argues that it is essential in strategic situations that players reason about conversations *as if* they were infinite, and models this via infinite games. However, all conversations are in fact finite; at some point the Jury just stops listening. To address this issue, [4] introduces a weighting function for the Jury that guarantees a winner in a finite number of moves. In order to preserve the strategic reasoning of players in infinitary situations, it is crucial that such weighting functions (or in this case the Jury winning conditions) not be known to the players. Here we satisfy this requirement by making the beliefs of the players about the Jury uncertain, in particular, the Jury winning conditions. However, we abstract out here the weighting function in the definition of the Jury winning condition itself (Definition 4) in the interests of simplicity.

Definition 6 (Utility). *In what follows, for every history* h *of an ME game, it will be convenient to assign a binary* utility $u_i(\mathsf{h})$ *for each player* i *defined as:*

$$u_i(\mathsf{h}) = \begin{cases} 1 & \text{if } \mathsf{h} \in Win_i \\ 0 & \text{otherwise} \end{cases}$$

Definition 7 (Pure strategy). *A pure strategy* σ_i *for Player* i *in an ME game is a function from the set of* $(1-i)$*-plays to moves in* V_i^+*. That is,* $\sigma_i : \mathcal{P}_{(1-i)} \to V_i^+$*. Let* S_i *denote the set of strategies for Player* i *and let* $S = S_0 \times S_1$.

[1] It is known that sigma algebras are not sufficient to reason about the information of the players while strategising and can lead to paradoxical results [10]. Such paradoxes are avoided in our setting because it is the Jury who determines the winning sets Win_0, Win_1 of the players and these sets are not subject to measurability restrictions.

Let $\rho = x_0 x_1 \ldots$ be a play in an ME game where $x_0 = \epsilon$ and let $\rho_j = x_0 x_1 \ldots x_j$ for $j > 0$ be the set of prefixes of ρ. We say that ρ conforms to a strategy σ_i of Player i if for every $(1-i)$-play ρ_j, $x_{j+1} = \sigma_i(\rho_j)$. Let ρ_{σ_i} denote the set of plays that conform to the strategy σ_i and let $\rho_{(\sigma_0,\sigma_1)}$ denote the *unique* play that conforms to the the strategy pair (σ_0, σ_1). Conversely, given a finite play ρ, we let S_i^ρ denote the set of all strategies of Player i such that ρ conforms to every strategy $\sigma_i \in S_i^\rho$ and let S^ρ denote the set of all strategy pairs such that ρ conforms to every $(\sigma_0, \sigma_1) \in S^\rho$.

To define a measure over S_i, note that S_i is a subset of $(V_i^+)^{\mathcal{P}_{(1-i)}}$ where $(V_i^+)^{\mathcal{P}_{(1-i)}}$ is the set of all functions from $\mathcal{P}_{(1-i)}$ to V_i^+. However, Aumann in [8] showed that even when both $\mathcal{P}_{(1-i)}$ are V_i^+ separable, not all subsets of $(V_i^+)^{\mathcal{P}_{(1-i)}}$ are measurable. Hence, from now on we shall only deal with measurable subsets of S_i and S. Moreover, when we say S_i or S we mean the largest measurable subsets of S_i and S respectively.

Example 4 *Let us now see how we can cast the court-room situation of Example 1 into an ME game. Let us label the conversation in terms of their* DU*s:*

a. **P:** $\langle \pi_1 : \phi_1 =$*Do you have any bank accounts in Swiss Banks, Mr. Bronston?*$\rangle$
c. **P:** $\langle \pi_2 : \phi_2 =$*Have you ever?*$\rangle$
e. **P:** $\langle \pi_3 : \phi_3 =$*Thank you Mr. Bronston.*$\rangle$
e'. **P:** $\langle \pi_3' : \phi_3' =$*Please do not try to mislead the Jury. Answer my question directly. I don't want to know if the company had an account at the Swiss banks. I want to know if you had an account there.*$\rangle$
b. **B:** $\langle \pi_4 : \phi_4 =$*No, sir.*$\rangle$
b'. **B:** $\langle \pi_4' : \phi_4' =$*Yes, sir.*$\rangle$
b. **B:** $\langle \pi_5 : \phi_5 =$*The company had an account there for about six months, in Zurich.*$\rangle$

We then formulate the ME game structure for the example together with continuations among the types we have mentioned as shown in Figure 2.

The dashed edges represent "all possible continuations". Note that after each history there might be infinitely many other moves by either player. However, these moves are irrelevant to the topic of the trial. Hence they are not depicted in our simplified picture. The labels on the edges represent the discourse relations between the DU s that are given by SDRT; for example Q-elab denotes a relation between an original question and a follow-up question, while IQAP denotes a relation between a question q and a DU *whose implicatures provide an answer to q. Similar to Example 2, the shaded region represents the resulting 5-turn play ρ_5 of Example 1 which is a* ULF *determined by the grammar. The* ULF *is a formula with a variable for the relation inferred between π_2 and π_5:*

$$\rho_5 = (\langle \pi_1 : \phi_1 \rangle \langle \pi_4 : \phi_4 \mathsf{qap}(\pi_1, \pi_4) \rangle \langle \pi_2 : \phi_2 \mathsf{qelab}(\pi_1, \pi_2) \rangle \langle \pi_5 : \phi_5 \exists \mathcal{R} \cdot \mathcal{R}(\pi_2, \pi_5) \rangle \langle \pi_3 : \phi_3 \mathsf{ack}(\pi_5, \pi_3) \rangle)$$

The relation between π_2 and π_5 is underspecified, because the grammar (syntax, compositional and lexical semantics) does not determine it. Each of the

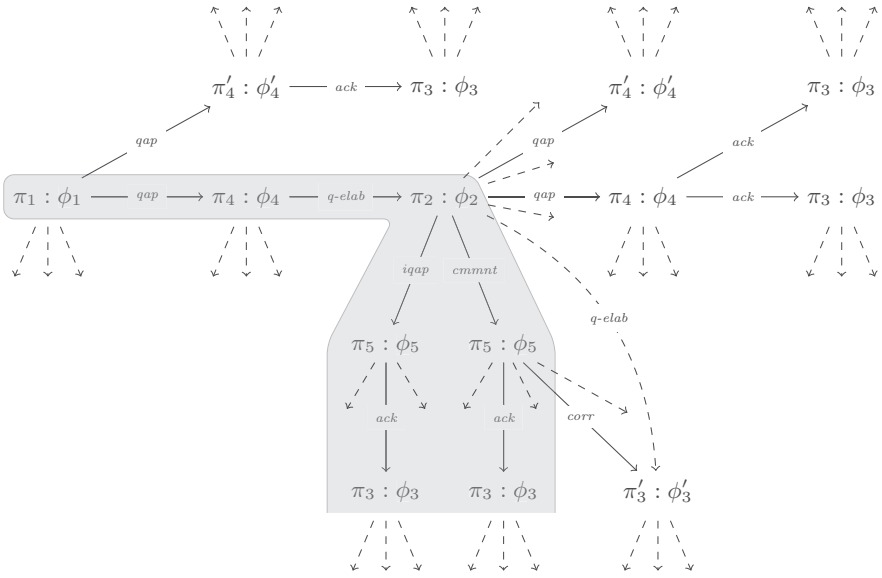

Fig. 2: The ME game tree for the conversation in Example 1

branches in the shaded region provides a different witness for $\mathcal{R}$ resulting in an SDRS (FLF). Each such SDRS is a history in the game. We consider only two such histories (SDRSs), $\mathsf{h}_{\mathsf{cmt}}, \mathsf{h}_{\mathsf{iqap}}$, in the present example to keep the presentation simple and uncluttered. These are the histories where $\mathcal{R}$ is interpreted as a 'comment' and an 'indirect question-answer pair' respectively. That is,

$$\mathsf{h}_{\mathsf{cmt}} = \langle \pi_1 : \phi_1 \rangle \langle \pi_4 : \phi_4 \mathsf{qap}(\pi_1, \pi_4) \rangle \langle \pi_2 : \phi_2 \mathsf{qelab}(\pi_1, \pi_2) \rangle \langle \pi_5 : \phi_5 \mathsf{cmt}(\pi_2, \pi_5) \rangle \langle \pi_3 : \phi_3 \mathsf{ack}(\pi_5, \pi_3) \rangle$$

$$\mathsf{h}_{\mathsf{iqap}} = \langle \pi_1 : \phi_1 \rangle \langle \pi_4 : \phi_4 \mathsf{qap}(\pi_1, \pi_4) \rangle \langle \pi_2 : \phi_2 \mathsf{qelab}(\pi_1, \pi_2) \rangle \langle \pi_5 : \phi_5 \mathsf{iqap}(\pi_2, \pi_5) \rangle \langle \pi_3 : \phi_3 \mathsf{ack}(\pi_5, \pi_3) \rangle$$

Many other histories could result from the ULF. For instance, $\mathcal{R}$ may also be interpreted as a 'background', a 'correction' etc. These witnesses are the result of uncertain inference that depends on several factors.[2]

Now, a history of the form

$$\mathsf{h}_y \in \begin{cases} \langle \pi_1 : \phi_1 \rangle \langle \pi_4' : \phi_4' \mathsf{qap}(\pi_1, \pi_4') \rangle \dots \\ \langle \pi_1' : \phi_1 \rangle \langle \pi_4 : \phi_4 \mathsf{qap}(\pi_1, \pi_4) \rangle \langle \pi_2 : \phi_2 \mathsf{qelab}(\pi_1, \pi_2) \rangle \langle \pi_4' : \phi_4' \mathsf{qap}(\pi_2, \pi_4') \rangle \dots \end{cases}$$

[2] [2] modelled the inference with a non-monotonic logic, with which we could infer the disjunction of the relations we have depicted above. But it could not do more than that. We assume here a more probabilistic inference relation to model implicatures, and below we will show how these probabilities are dependent on beliefs about interlocutors.

represents a sequence where B has responded to P's question with a Yes. These are direct answers, and given our assumptions in the introduction about P's winning conditions encoded by the Jury, such histories are losing for B. Also, as no continuation of such a history is relevant to the outcome of the trial, such plays are finitely decided. Similarly, a history where B has responded to P's question with a No is also losing for B given our assumptions and are finitely decided.

The most interesting histories are $\mathsf{h}_{\mathsf{cmt}}$ *and* $\mathsf{h}_{\mathsf{iqap}}$ *which are the ones generated from the play which actually transpired in the court-room scene. Whether the Jury interprets* ρ_5 *as* $\mathsf{h}_{\mathsf{cmt}}$ *or as* $\mathsf{h}_{\mathsf{iqap}}$ *and whether such a history is winning for B or not depends on the nature of the Jury, or its 'type'. We look at two cases:*

Case(i): the case where the Jury interprets ρ_5 *as* $\mathsf{h}_{\mathsf{iqap}}$ *and puts* $\mathsf{h}_{\mathsf{iqap}} \in Win_P$*. Note, that this is the case where the Jury interprets the move* π_5 *by B as having the implicature:*
(I) The company had an account in the Swiss banks **but** *Bronston himself did not.*

Case(ii): the case where the Jury interprets ρ_5 *as* $\mathsf{h}_{\mathsf{cmt}}$ *and puts* $\mathsf{h}_{\mathsf{cmt}} \in Win_B$ *for the lack of a follow-up from P.*

As conversation is supposedly a rational activity, we should try to determine whether B is rational in playing $\pi_5 \exists \mathcal{R}.\mathcal{R}(\pi_2, \pi_5)$ *after* π_2*? What about P? Is he rational when he plays* $\pi_3 \mathsf{ack}(\pi_5, \pi_3)$ *after* π_5*? Intuition says that in case (i) P was rational and so was B, whereas in (ii) P was not playing optimally. And we shall see that the Jury, or the players' estimates of the Jury, plays a crucial role in deciding rational play.*

To flesh this out, we turn to the epistemic structure of the above ME game.

4 The dynamics of ME games

In this section we describe the dynamics of ME games. We study ME games from the perspectives of the players, the Jury, and a third party observer (us). What do the players think and believe as they make their moves? Of course, every player wants to 'win', to achieve what he believes is the winning goal set for him by the Jury. Based on this belief and also what he believes of the other players' approach, he strategizes for every next move.

4.1 Types, beliefs and interpretations

The type of a player i is an abstract object that is supposed to code-up anything and everything about the player, including his behaviour, the way he strategizes, his personal biases, etc. [12].

Definition 8 (Harsanyi type space [12]). *A Harsanyi* type space *for S is a tuple* $\mathcal{T} = (\{T_i\}_{i\in\{0,1\}}, T_{\mathcal{J}}, \{\hat{\beta}_i^{\rho}\}_{i\in\{0,1\},\rho\in\mathcal{P}}, \{\hat{\beta}_{\mathcal{J}}^{\rho}\}_{\rho\in\mathcal{P}}, S)$ *such that* $T_{\mathcal{J}}$ *and* T_i*, for each i, are non-empty (at-most countable) sets called the* Jury-types *and i-*types *respectively and* $\{\hat{\beta}_i^{\rho}\}$ *and* $\{\hat{\beta}_{\mathcal{J}}^{\rho}\}$ *are the beliefs of Player i and the Jury respectively at play* $\rho \in \mathcal{P}$ *and are defined below.*

We are interested in the beliefs of the players about other players and about what they have said and how these two things influence each other. So we will separate out the effect of types both on beliefs about other players and on interpretations of a conversation that result in particular histories.

Definition 9 (Belief function). *For every play $\rho \in \mathcal{P}$ the* (first order) belief $\hat{\beta}_i^\rho$ *of player i at ρ is a pair of measurable functions $\hat{\beta}_i^\rho = (\beta_i^\rho, \xi_i^\rho)$ where β_i^ρ is the* belief function *and ξ_i^ρ is the* interpretation function *defined as:*

$$\beta_i^\rho : T_i \times \mathsf{h}(\rho) \to \Delta(T_{(1-i)} \times S^\rho_{(1-i)} \times T_{\mathcal{J}})$$

$$\xi_i^\rho : T_i \times T_{(1-i)} \times T_{\mathcal{J}} \to \Delta(\mathsf{h}(\rho))$$

Similarly the (first order) belief $\hat{\beta}^\rho_{\mathcal{J}}$ of the Jury is a pair of measurable functions $\hat{\beta}^\rho_{\mathcal{J}} = (\beta^\rho_{\mathcal{J}}, \xi^\rho_{\mathcal{J}})$ where the belief function *β_i^ρ and the* interpretation function *ξ_i^ρ are defined as:*

$$\beta^\rho_{\mathcal{J}} : T_{\mathcal{J}} \times \mathsf{h}(\rho) \to \Delta(T_0 \times S_0^\rho \times T_1 \times S_1^\rho)$$

$$\xi^\rho_{\mathcal{J}} : T_{\mathcal{J}} \times T_0 \times T_1 \to \Delta(\mathsf{h}(\rho))$$

Note that $\mathcal{T}_i$ and $\mathcal{T}_{\mathcal{J}}$ are equipped with the discrete topology and $(T_{(1-i)} \times S^\rho_{(1-i)} \times T_{\mathcal{J}})$ the product topology as usual and receives the product measure. Hence the belief functions can also be written as:

$$\beta_i^\rho : T_i \times \mathsf{h}(\rho) \to \Delta(T_{(1-i)}) \times \Delta(S^\rho_{(1-i)}) \times \Delta(T_{\mathcal{J}})$$

$$\beta^\rho_{\mathcal{J}} : T_{\mathcal{J}} \times \mathsf{h}(\rho) \to \Delta(T_0) \times \Delta(S_0^\rho) \times \Delta(T_1) \times \Delta(S_1^\rho)$$

Intuitively, for any type of the player or the Jury, the respective interpretation function says how they interpret the current play as; that is, what are the probabilities that they assign to each possible history arising from the current play. The belief function then gives their beliefs about the types and the strategies of the other players and/or the Jury for this interpretation. That the interpretation function returns a probability distribution over histories is consonant with the way computational linguists like [1] model how various features of the play lead to a probability distribution over full SDRSs, something we mentioned in the previous section.

For certain examples, the beliefs or the interpretations of the players or the Jury may be independent of one or more components.[3] In that case we shall simply suppress those components. For example, the winning condition for the players might depend entirely on the Jury's belief about the type of just one of the players, Player i (say) and be independent of what it believes about the type of Player $(1-i)$ or the strategies of the players. In that case the belief of the Jury is given by the function $\beta^\rho_{\mathcal{J}} : T_{\mathcal{J}} \times \mathsf{h}(\rho) \to \Delta(T_i)$ Such independence will often simplify our analyses of some examples below.

[3] A function $f : A_1 \times A_2 \times \ldots A_n \to B$ is independent of the jth component, $1 \leq j \leq n$, if for all $a_j, a'_j \in A_j$, $f(a_1, a_2, \ldots, a_j, \ldots, a_n) = f(a_1, a_2, \ldots, a'_j, \ldots, a_n)$.

We will also often deal with these functions when one or more of the components are *fixed*. For example, suppose we want to talk about the interpretation of a type of the Jury given that the types of the players are t_0 and t_1 respectively. Then, we talk about the interpretation function of the Jury restricted to types t_0 and t_1 and is given as $\xi^{\rho}_{\mathcal{J}}|_{(t_0 \times t_1)} : T_{\mathcal{J}} \times \{t_0\} \times \{t_1\} \rightarrow \Delta(\mathsf{h}(\rho))$ When the independent component(s) are clear from the context, we shall often supress them in the description of the functions. The above function would then be given simply as: $\xi^{\rho}_{\mathcal{J}} : T_{\mathcal{J}} \rightarrow \Delta(\mathsf{h}(\rho))$

S, the set of strategies over linguistic moves in ME games, is an uncountable set. Thus, the measures provided by belief functions are probability density functions over subsets in $(T_{(1-i)} \times S_{(1-i)} \times T_{\mathcal{J}})$, something we exploit to define higher order beliefs, beliefs that players or the Jury have about the beliefs of other players (and the Jury) (for details see the Appendix).

4.2 The Bronston example revisited

Example 5 *In light of these epistemic notions for ME games, let us revisit our example of Bronston and the Prosecutor. A strategy is said to be consistent (coherent) if all plays conforming to it are linguistically consistent (coherent) [see [6] for a formal definition of coherence and consistency]. After his elaboration question π_2, we can classify the coherent and consistent strategies of P into two categories and those of B into three. We formulate these strategies in terms of of the classes of moves as discussed at the end of Section 3.1.*

- *Strategies where P plays $\pi_3\mathsf{ack}(\pi_\alpha, \pi_3)$, $\pi_\alpha \in \{\pi_4, \pi'_4, \pi_5\}$ on his turn after B plays any move of the form (class) π_4, π'_4, or π_5. Crucially on this strategy P acknowledges B's indirect answer. We denote such a kind of strategy as σ_{ack}.*
- *Strategies where P plays $\pi'_3\mathsf{qelab}(\pi_4, \pi'_3)$, that is, asks a follow-up question on some turn after B has played π_4 and plays $\pi_2\mathsf{ack}(\pi, \pi_2)$, $\pi \in \{\pi_4, \pi'_4\}$ in response to π_4 and π'_4. We denote such a strategy as σ_{qelab}.*

Here is the specification of B's strategies:

- *Strategies where he eventually admits a yes. Such a strategy will be denoted as τ_{yes}.*
- *Strategies where he eventually admits a no. Such a strategy will be denoted as τ_{no}.*
- *Finally, strategies where he sticks to something of the form π_5. Such a strategy will be denoted as τ_{ind}.*

Thus, after the question π_2 by P, we can partition B's strategies of interest into three sets, and we can consider $S_B = \{\tau_{\mathsf{yes}}, \tau_{\mathsf{no}}, \tau_{\mathsf{ind}}\}$. Similarly, after B's moves, P has effectively two strategies $S_P = \{\sigma_{\mathsf{ack}}, \sigma_{\mathsf{qelab}}\}$. Thus, we can just consider S for our example as a finite set: $S = S_P \times S_B$. Let us suppose that P and B each have two types: $T_P = \{t_P, t'_P\}$, and $T_B = \{t_B, t'_B\}$. We will also

not consider types of the Jury in this example; so $T_{\mathcal{J}} = \emptyset$. *Let* ρ_3 *be the play:* $\rho_3 = \pi_1\pi_4\mathsf{qap}(\pi_1, \pi_4)\pi_2\mathsf{qelab}(\pi_4, \pi_2)$. ρ_3 *in itself is a complete SDRS and hence* $\mathsf{h}(\rho_3)$ *contains only one history* h *(say). Thus, the interpretation functions* $\xi_B^{\rho_3}$ *and* $\xi_P^{\rho_3}$ *for* B *and* P *both trivially map* ρ_3 *to* h *irrespective of their types. That is, all the types of both the players are sure about the history being* h.

Now suppose, after the play ρ_3, *the (first-order) belief maps for each of these types are described by the tables 1a and 1b.*

$\beta_P^{\rho_3}(t_P, \mathsf{h})$	τ_{yes}	τ_{no}	τ_{ind}
t_B	0.5	0.3	0
t'_B	0	0	0.2

$\beta_P^{\rho_3}(t'_P, \mathsf{h})$	τ_{yes}	τ_{no}	τ_{ind}
t_B	0.3	0.2	0
t'_B	0	0	0.5

Table 1a: P's beliefs about B after the play ρ_3

Table 1a says that after ρ_3, *type* t_P *of* P *believes that* B *is of type* t_B *and plays a strategy of the form* τ_{yes} *with probability 0.5, of the form* τ_{no} *with probability 0.3 and of the form* τ_{ind} *with probability 0.* t_P *also believes that* B *is of type* t'_B *and plays a strategy of the form* τ_{yes} *with probability 0, of the form* τ_{no} *with probability 0 and of the form* τ_{ind} *with probability 0.2. Note that the individual probabilities sum to 1 as should be the case. Table 1a also says that after* ρ_3, *type* t'_P *of* P *believes that* B *is of type* t_B *and plays a strategy of the form* τ_{yes} *with probability 0.3, of the form* τ_{no} *with probability 0.2 and of the form* τ_{ind} *with probability 0.* t'_P *also believes that* B *is of type* t'_B *and plays a strategy of the form* τ_{yes} *with probability 0, of the form* τ_{no} *with probability 0 and of the form* τ_{ind} *with probability 0.5. Table 1b is similar.*

We can now calculate the first and higher order beliefs of P *and* B. *Both* P *and* B *have two types. If* B *is of the type* t_B *then he he assigns an equal probability to* P *being of type* t_P *and* t'_P *and equal probabilities to* P*'s playing a strategy of the form* σ_{ack} *or* σ_{qelab}. *On the other hand if he is of type* t'_B, *although he assigns both of* t_P *and* t'_P *equal probability, he is certain that* P *plays a strategy of the form* σ_{qelab} *in both cases. Next, type* t_P *of* P *assigns a probability of 0.8 to the type* t_B *and a probability of 0.2 to the type* t'_B *of* B *respectively (as can be seen by summing the rows for each type in Table 1a). Thus type* t_P *of* P *believes that it is likely with probability 0.8 that (i)* B *thinks there is a 50% chance that he plays a strategy of the form* σ_{ack} *and likely with the remaining probability 0.2 that (ii)* B *thinks for certain that he plays a strategy of the form* σ_{qelab}. *And similarly for type* t'_P *of* P. *These are second-order beliefs of* P. *We can continue this way for the higher order beliefs of* P *and* B.

Now, B *plays* $\pi_5\exists\mathcal{R} \cdot \mathcal{R}(\pi_2, \pi_5)$ *and the play* ρ_3 *is extended to* ρ_4 *as:*

$$\rho_4 = \pi_1\pi_4\mathsf{qap}(\pi_1, \pi_4)\pi_2\mathsf{qelab}(\pi_4, \pi_2)\pi_5\exists\mathcal{R} \cdot \mathcal{R}(\pi_2, \pi_5)$$

As we saw in Eg. 4, there are (at-least) two relevant histories (SDRSs) in the set $\mathsf{h}(\rho_4)$. *These are depicted in the shaded region of Figure 2 and are given as follows:*

$\beta_B^{\rho_3}(t_B, \mathsf{h})$	σ_{ack}	σ_{qelab}
t_P	0.5	0
t'_P	0	0.5

$\beta_B^{\rho_3}(t'_B, \mathsf{h})$	σ_{ack}	σ_{qelab}
t_P	0	0.5
t'_P	0	0.5

Table 1b: B's beliefs about P after the play ρ_3

$$\mathsf{h}_{\mathsf{cmt}} = \pi_1\pi_4\mathsf{qap}(\pi_1, \pi_4)\pi_2\mathsf{qelab}(\pi_4, \pi_2)\pi_5\mathsf{cmt}(\pi_2, \pi_5)$$

$$\mathsf{h}_{\mathsf{iqap}} = \pi_1\pi_4\mathsf{qap}(\pi_1, \pi_4)\pi_2\mathsf{qelab}(\pi_4, \pi_2)\pi_5\mathsf{iqap}(\pi_2, \pi_5)$$

B and P interpret ρ_4 in different ways. In ρ_4 t_P and t'_P are confronted with an indirect response. This will affect their interpretations. t_P might be more likely to take this indirect response as an IQAP and so $\xi_P^{\rho_4}(t_P)(\mathsf{h}_{\mathsf{iqap}}) = 0.7, \xi_P^{\rho_4}(t_P)(\mathsf{h}_{\mathsf{cmt}}) = 0.3$. while the interpretations of t'_P, who was more attuned to the possibility of an indirect response and might be more suspicious, after the play ρ_4 are: $\xi_P^{\rho_4}(t'_P)(\mathsf{h}_{\mathsf{cmt}}) = \xi_P^{\rho_4}(t'_P)(\mathsf{h}_{\mathsf{iqap}}) = 0.5$.

Furthermore, suppose irrespective of his type B did intend the move as an indirect response to P's question. In that case, the the interpretation function of t_B and t'_B after the play ρ_4 can be given as: $\xi_B^{\rho_4}(t_B)(\mathsf{h}_{\mathsf{cmt}}) = 0$ and $\xi_B^{\rho_4}(t'_B)(\mathsf{h}_{\mathsf{iqap}}) = 1$.

In general, the two histories may affect B and P's beliefs about the type-strategy pairs of the other player – $\mathsf{h}_{\mathsf{cmt}}$ and $\mathsf{h}_{\mathsf{iqap}}$ may keep the probabilities unchanged for t_P while they may change the probability distribution for t'_P over B's strategies; he might assign τ_{ind} a higher probability for t'_B. This in turn might shift B's beliefs – t'_B might assign a higher probability to t'_P's follow up strategy. We shall explore such dependence of the beliefs of the players on the interpreted histories, in our extended example in Section 4.5. For the current example, we assume that the beliefs of P and B are independent of their interpretation of the current history at ρ_4. They are then given in tables 2a and 2b.

$\beta_P^{\rho_4}(t_P)$	τ_{yes}	τ_{no}	τ_{ind}
t_B	0.5	0.3	0
t'_B	0	0	0.2

$\beta_P^{\rho_4}(t'_P)$	τ_{yes}	τ_{no}	τ_{ind}
t_B	0.3	0.1	0
t'_B	0	0	0.6

Table 2a: P's beliefs about B after the play ρ_4

A last piece of business in our basic epistemic set up is this. Recall that the Jury winning condition Win_i for each player i was defined as a subset of the set of all possible histories of the ME game. Now the interpretations every type of each player has on the set of possible histories for a play of the ME game naturally leads to expected utilities, which we define in the Appendix.

$\beta_B^{\rho_4}(t_B)$	σ_{ack}	σ_{qelab}
t_P	0.5	0
t'_P	0	0.5

$\beta_B^{\rho_4}(t'_B)$	σ_{ack}	σ_{qelab}
t_P	0	0.5
t'_P	0	0.5

Table 2b: B's beliefs about P after the play ρ_4

4.3 Rationality and common belief in rationality

With the belief structures of our players now set, we can define rationality and common belief in rationality. Given a play ρ let $E^\rho \subset (T_0 \times S_0^\rho \times T_1 \times S_1^\rho \times T_\mathcal{J})$. Let

$$E_i^\rho = \{(t_{(1-i)}, \sigma_{(1-i)}, t_\mathcal{J}) \mid (t_0, \sigma_0, t_1, \sigma_1, t_\mathcal{J}) \in E^\rho\}$$

and

$$E_\mathcal{J}^\rho = \{(t_0, \sigma_0, t_1, \sigma_1) \mid (t_0, \sigma_0, t_1, \sigma_1, t_\mathcal{J}) \in E^\rho\}$$

be the events at ρ for Player i and the Jury respectively. Let $B(E^\rho) \subset (T_0 \times S_0^\rho \times T_1 \times S_1^\rho \times T_\mathcal{J})$ be defined as

$$B(E^\rho) = \{(t_0, \sigma_0, t_1, \sigma_1, t_\mathcal{J}) \mid (t_{(1-i)}, \sigma_{(1-i)}, t_\mathcal{J}) \in B_i(E_i^\rho) \text{ and } (t_0, \sigma_0, t_1, \sigma_1) \in B_\mathcal{J}(E_\mathcal{J}^\rho)\}$$

where $B_i(E_i^\rho)$ and $B_\mathcal{J}(E_\mathcal{J}^\rho)$ are the type-strategy pairs that believe in E_i^ρ and $E_\mathcal{J}^\rho$ respectively as defined in the Appendix. Then 'E^ρ and common belief that E^ρ at ρ' is defined as the event that E^ρ and everyone believes that E^ρ at ρ and everyone believes at ρ that everyone believes that E^ρ at ρ and so on. More formally,

Definition 10 (Common belief). *Let*

$$B^0(E^\rho) = E^\rho \text{ and for } n \geq 1, B^n(E^\rho) = B(B^{n-1}(E^\rho))$$

Then E^ρ and common belief of E^ρ at ρ *is defined as the following infinite conjunction:*

$$C(E^\rho) = \bigcap_{n \geq 0} B^n(E^\rho)$$

We can then define expected payoffs of a strategy for a given type of player and an appropriate probability distribution p over strategies (see the Appendix), and henceforth talk of a strategy that maximizes an expected payoff or is a *best response* for a particular type of player with respect to p. We use these concepts to define:

Definition 11 (Rationality and common belief in rationality). *For a play ρ, let $R[\rho]_i$ be defined as:*

$$R[\rho]_i = \{(t_i, \sigma_i, t_\mathcal{J}) \mid t_i \in T_i, t_\mathcal{J} \in T_\mathcal{J} \text{ and } \sigma_i \text{ is a best response to } p_{t_i}^\rho \text{ for } t_i\}$$

where $p_{t_i}^\rho$ is the probability measure over $S_{(1-i)}$ generated by t_i as defined in the Appendix.

These are the type-strategy pairs of Player i where she plays rationally *at ρ. Now, let*

$$R[\rho] = \{(t_0, \sigma_0, t_1, \sigma_1, t_{\mathcal{J}}) \mid (t_0, \sigma_0, t_{\mathcal{J}}) \in R[\rho]_0 \text{ and } (t_1, \sigma_1, t_{\mathcal{J}}) \in R[\rho]_1\}$$

$R[\rho]$ is the event that both players play rationally at ρ. Then

$$\mathsf{RCBR}[\rho] = C(R[\rho])$$

is the event where Rationality and Common Belief in Rationality *holds at ρ.*

Thus $\mathsf{RCBR}[\rho]$ are the type-strategy tuples where both the players are rational, each believes that the other is rational, each believes that the other believes that the other is rational and so on. In addition, the Jury believes that the players are rational, the players believe that the Jury believes that they are rational and so on.

Let us now see how RCBR applies to our Example 5, as we have so far developed it.

Example 6 *Recall from Example 5 that after the play*

$$\rho_3 = \pi_1\pi_4\mathsf{qap}(\pi_1, \pi_4)\pi_2\mathsf{qelab}(\pi_4, \pi_2)$$

type t_P of P assigns a probability of 0.5 that B is of type t_B and plays a strategy of the form τ_{yes}, a probability of 0.3 that B is of type t_B and plays a strategy of the form τ_{no}, and a probability of 0.2 that B is of type t'_B and he plays a strategy of the form τ_{ind}. We also saw in Example 5 that after observing the move $\pi_5\exists\mathcal{R} \cdot \mathcal{R}(\pi_2, \pi_5)$ by B, which results in the play ρ_4, although he forms different interpretations about the possible histories generated by the move, t_P's beliefs about the type-strategy pairs of B do not change.

Let us now look at the two cases envisaged in Example 4, which depended on how the Jury interprets the play between P and B.

Case (i): *whether P acknowledges B's indirect response or not, the play remains within Win_P. Thus, given our assumptions about possible moves in this game, both ρ_3 and ρ_4 are in Win_P. Then both σ_{ack} and σ_{qelab} are optimal for t_P.*

Now, since ρ_3 is an FLF*, let $\mathsf{h}(\rho_3)$ be the singleton set $\{\mathsf{h}\}$. Since we are ignoring the types of the Jury for now and since $\rho_3 \in Win_P$, we have*

$$R[\rho_3]_P = \{(t_P, \sigma_{\mathsf{ack}}), (t_P, \sigma_{\mathsf{qelab}}), (t'_P, \sigma_{\mathsf{ack}}), (t'_P, \sigma_{\mathsf{qelab}})\}$$

Once P updates his beliefs after observing $\pi_5\exists\mathcal{R} \cdot \mathcal{R}(\pi_2, \pi_5)$, the set of optimal strategies for him does not change (since $\rho_4 \in Win_P$ as well). Thus

$$R[\rho_4]_P = \{(t_P, \sigma_{\mathsf{ack}}), (t_P, \sigma_{\mathsf{qelab}}), (t'_P, \sigma_{\mathsf{ack}}), (t'_P, \sigma_{\mathsf{qelab}})\} = R[\rho_3]_P$$

Similarly, for B,

$$R[\rho_4]_B = \{(t_B, \tau_{\mathsf{yes}}), (t_B, \tau_{\mathsf{no}}), (t_B, \tau_{\mathsf{ind}})(t'_B, \tau_{\mathsf{yes}}), (t'_B, \tau_{\mathsf{no}}), (t'_B, \tau_{\mathsf{ind}})\}$$

since in case (i), every play in the ME game is losing for B, either immediately or eventually. So, $R[\rho_4] = R[\rho_4]_P \times R[\rho_4]_B$. *Now,* $B_P(R[\rho_4]_B) = R[\rho_4]_P$ *and* $B_B(R[\rho_4]_P) = R[\rho_4]_B$. *Thus* $B(R[\rho_4]) = B_P(R[\rho_4]_B) \times B_B(R[\rho_4]_P) = R[\rho_4]$. *Hence* RCBR *holds for all type-strategy pairs in this case, given how we have assigned the utilities. Note that P's and B's beliefs about each other do not matter here.*

Case (ii): *This is the more interesting case, where an acknowledgement of* π_5 *by P potentially moves the conversation out of* Win_P. *That is to say,* $\rho_5 = \rho_4\pi_3\mathsf{ack}(\pi_5, \pi_3) \notin Win_P$. *In this case, once P updates with the observation of* $\pi_5 \exists \mathcal{R} \cdot \mathcal{R}(\pi_2, \pi_5)$, *the acknowledgement* σ_{ack} *is not an optimal follow-up for him. Hence,* $R[\rho_4]_P = \{(t_P, \sigma_{\mathsf{qelab}}), (t'_P, \sigma_{\mathsf{qelab}})\}$. *For B, things are more complicated. The expected utility of* σ_{ind} *for* t_B *is greater than that for his other responses, because if P is of type* t_P, *which B estimates as having a probability of 0.5, the conversation has some chance of becoming winning for B and hence have a non-zero expected utility, whereas his other responses will net him a certain loss and hence 0 utility. On the other hand, given that a follow up question will inevitably follow for* t'_B, *any of B's options are optimal. Thus,* $R[\rho_4]_B = \{(t_B, \tau_{\mathsf{ind}}), (t'_B, \tau_{\mathsf{yes}}), (t'_B, \tau_{\mathsf{no}}), (t'_B, \tau_{\mathsf{ind}})\}$ *and* $R[\rho_4] = R[\rho_4]_P \times R[\rho_4]_B$.

Next, from the tables 2a and 2b, we have that $B_P(R[\rho_4]_B) = \{(t_P, \sigma_{\mathsf{qelab}}), (t'_P, \sigma_{\mathsf{qelab}})\} = R[\rho_4]_P$ *and* $B_B(R[\rho_4]_P) = \{(t'_B, \tau_{\mathsf{yes}}), (t'_B, \tau_{\mathsf{no}}), (t'_B, \tau_{\mathsf{ind}})\}$ *and* $B(R[\rho_4]) = B_B(R[\rho_4]_P) \times B_P(R[\rho_4]_B)$. *In the next iteration, we have* $B_B(B_P(R[\rho_4]_B)) = B_B(R[\rho_4]_P)$ *and* $B_P(B_B(R[\rho_4]_P)) = B_P(R[\rho_4]_B)$ *and we have reached a fixed point. Thus we have,* $\mathsf{RCBR}[\rho_4] = B(R[\rho_4]) = \{(t_P, \sigma_{\mathsf{qelab}}), (t'_P, \sigma_{\mathsf{qelab}})\} \times \{(t'_B, \tau_{\mathsf{yes}}), (t'_B, \tau_{\mathsf{no}}), (t'_B, \tau_{\mathsf{ind}})\}$.

This accords with intuitions, as in this case it is always optimal for P to ask a follow-up question to B's indirect response π_5. *Type* t'_B *of B is the only type that is sure about this and hence whatever he plays will be losing for him.*

4.4 Jury types

The epistemic picture for our original example 1 is still incomplete. In Example 6, type t_P of P is not rational, if, as in case (ii), $\rho_5 \notin Win_P$. P actually pursued σ_{ack} in real life. So manifestly for him, case (i) was what the Jury did. But what about B? Intuitively, the indirect strategy seems better than the direct answers, at least in hindsight. Our analysis so far does not capture that. However, to correctly formulate our intuitions, we need to include in our analysis, types for the Jury and the players' estimation of these types. It is the different Jury types that assign different winning conditions to the game and that is what ultimately dictates rational conversational behavior or not. And this is incidentally one reason why postulating a Jury is an essential feature of an ME game.

We now complete our picture by introducing types for the Jury in our running courtroom example conversation between Bronston and the Prosecutor. This allows us to perform a complete analysis and reflect on whether Bronston was indeed rational in his indirect response to the Prosecutor's question.

Example 7 *We preserve the same setting as in Example 6 but now introduce types for the Jury which were missing so far. Let us first look at the possible types*

$\xi_{\mathcal{J}}^{\rho_4}(\ \cdot\ )$	$\mathsf{h}_{\mathsf{cmt}}$	$\mathsf{h}_{\mathsf{iqap}}$	
tj_1	0	1	$\in Win_P$
tj_2	1	0	$\in Win_P$
tj_3	0	1	$\notin Win_P$

Table 3a: Types of the Jury and its interpretation structure

for the Jury. We will assume that all Jury types assign all plays conforming to B's strategies τ_{yes} *and* τ_{no} *to* Win_P.

The Jury types differ, however, in how they treat B's response π_5. *As stated before,* π_5 *has the linguistic implicature (I).*

(I) The company had an account in the Swiss banks **but** Bronston himself did not.

We also assumed that B's move π_5 *gives rise to (at least) two histories* $\mathsf{h}_{\mathsf{cmt}}$ *and* $\mathsf{h}_{\mathsf{iqap}}$. $\mathsf{h}_{\mathsf{iqap}}$ *is the history in which B's response* π_5 *is treated as implying (I). Now based on whether the Jury interprets B's response as* $\mathsf{h}_{\mathsf{cmt}}$ *or* $\mathsf{h}_{\mathsf{iqap}}$ *and whether it treats such an interpretation as winning or losing for B, we can formulate at least three relevant types for the Jury.*

tj_1*: is the type of the Jury who interprets B's response* π_5 *as (I), treats the play* ρ_4 *as* $\mathsf{h}_{\mathsf{iqap}}$ *and puts it in the winning set for P,* Win_P. *This type exemplifies the Jury in our case (i) from Example 4.*

tj_2*: is the type of the Jury who does not accept implicature (I) and interprets* ρ_4 *as* $\mathsf{h}_{\mathsf{cmt}}$. *Such a type takes B not to have answered P's question and puts* $\mathsf{h}_{\mathsf{cmt}} \in Win_P$.

tj_3*: is the type of the Jury who interprets* ρ_4 *as* $\mathsf{h}_{\mathsf{iqap}}$ *but crucially expects more evidence in (perhaps) the way of more interrogation by the Prosecutor and hence puts* $\mathsf{h}_{\mathsf{iqap}} \notin Win_P$. *This type exemplifies the Jury in our case (ii).*

All these types of the Jury determine winning conditions for the players based solely on the histories corresponding to the play, which they interpret independent of their beliefs about the types of the players. Given this, we represent the interpretations of the Jury types for the play ρ_4 *with Table 3a.*

To consider relevant types for B and P, it is perhaps most intuitive to base them on their beliefs about the type of the Jury and what they believe about the beliefs of the opponent about the Jury type. Also, to simplify the analysis, we consider belief functions that assign probabilities of 0 or 1 only. We further assume that the beliefs of B are independent of P's strategies and also of the current history. Under these assumptions, consider 3 types each of B and P, denoted as $\mathsf{tb}_1, \mathsf{tb}_2, \mathsf{tb}_3$ *and* $\mathsf{tp}_1, \mathsf{tp}_2, \mathsf{tp}_3$ *respectively, such that,* tb_1 *and* tp_1 *believe that the Jury is of type* tj_1, tb_2 *and* tp_2 *believe that the Jury is of type* tj_2 *and* tb_3 *and* tp_3 *believe that the Jury is of type* tj_3. *These beliefs can be represented compactly as shown in left Table 3b(i). The first row of that table says that type* tb_1 *of B believes that the Jury is of type* tj_1 *and also believes that P believes that*

$\beta_B^{\rho_4}(\,\cdot\,)$	tp_1	tp_2	tp_3	tj_1	tj_2	tj_3
tb_1	0	1	0	1	0	0
tb_2	0	0	1	0	1	0
tb_3	1	0	0	0	0	1

$\xi_B^{\rho_4}(\,\cdot\,)$	$\mathsf{h}_{\mathsf{cmt}}$	$\mathsf{h}_{\mathsf{iqap}}$
tb_1	0	1
tb_2	0	1
tb_3	0	1

Table 3b(i): B's beliefs about the types of P and the Jury and his interpretations of the play ρ_4

the Jury, on the contrary, is of type tj_2*. The rows give other options understood similarly.*

It is also intuitive to assume that B indeed intended his move π_5 *as an indirect response to P's question. Hence, every type of B interprets the play* ρ_4 *as* $\mathsf{h}_{\mathsf{iqap}}$ *and this interpretation is independent of the type of P, as represented in the right Table 3b(i).*

Given our assumptions about the Jury, any strategy of the form τ_{yes} *or* τ_{no} *is losing for B. Hence, irrespective of the type t of B and his beliefs, the type-strategy pair* (t, τ_{ind}) *is always rational for him. And if he assigns a non-zero probability to the case that the Jury is actually of type* tj_3*, he stands a chance of being acquitted. Thus, when B responds with the indirect answer* π_5 *to P's question, he is indeed being rational and optimal.*

Next we consider three types for P. We shall look at the beliefs of P after the play ρ_4*. As in the case of the beliefs of B, we shall assume that the beliefs of P are independent of the strategies of B. However, at each of the histories* $\mathsf{h}_{\mathsf{cmt}}$ *and* $\mathsf{h}_{\mathsf{iqap}}$*, P's beliefs are consistent with what he beliefs about the type of the Jury, which recall were based in turn on the Jury's interpretation of the the play* ρ_4*. These beliefs functions are then given by Table 3c(i). Table 3c(i) is read just as Table 3b(i).*

$\beta_P^{\rho_4}(\cdot, \mathsf{h}_{\mathsf{cmt}})$	tb_1	tb_2	tb_3	tj_1	tj_2	tj_3
tp_1	1	0	0	0	1	0
tp_2	1	0	0	0	1	0
tp_3	0	0	1	0	1	0

$\beta_P^{\rho_4}(\cdot, \mathsf{h}_{\mathsf{iqap}})$	tb_1	tb_2	tb_3	tj_1	tj_2	tj_3
tp_1	1	0	0	1	0	0
tp_2	1	0	0	1	0	0
tp_3	0	0	1	0	0	1

Table 3c(i): P's beliefs about the types of B and the Jury

For the interpretations of ρ_4 *by P, we assume that they are in agreement with what each type of P believes about the interpretation of* ρ_4 *by the Jury and are independent of the type of B. Then, the interpretation function of P for* ρ_4 *is given as in Table 3c(ii).*

So, among these types and beliefs, which type-strategy pairs of P are rational? There are 5 such pairs:

$\xi_P^{\rho_4}(\ \cdot\ , \mathsf{tj}_1)$	$\mathsf{h}_{\mathsf{cmt}}$	$\mathsf{h}_{\mathsf{iqap}}$
tp_1	0	1
tp_2	0	1
tp_3	0	1

$\xi_P^{\rho_4}(\ \cdot\ , \mathsf{tj}_2)$	$\mathsf{h}_{\mathsf{cmt}}$	$\mathsf{h}_{\mathsf{iqap}}$
tp_1	1	0
tp_2	1	0
tp_3	1	0

$\xi_P^{\rho_4}(\ \cdot\ , \mathsf{tj}_3)$	$\mathsf{h}_{\mathsf{cmt}}$	$\mathsf{h}_{\mathsf{iqap}}$
tp_1	0	1
tp_2	0	1
tp_3	0	1

Table 3c(ii): P's interpretations of the play ρ_4

- $(\mathsf{tp}_1, \sigma_{\mathsf{ack}})$ *and* $(\mathsf{tp}_1, \sigma_{\mathsf{qelab}})$*: Indeed, because P himself interprets the current history as* $\mathsf{h}_{\mathsf{iqap}}$ *and believes that the Jury does the same, accepting the obvious implicature (I) of B's response* π_5 *making continuations of such plays winning for P. Hence he can either go on, having been assured of a commitment, or ask a follow up question.*
- $(\mathsf{tp}_2, \sigma_{\mathsf{ack}})$ *and* $(\mathsf{tp}_2, \sigma_{\mathsf{qelab}})$*: Since P himself interprets the current history as* $\mathsf{h}_{\mathsf{cmt}}$ *and also believes that the Jury does not take on implicature (I) of B's response interpreting the current history as* $\mathsf{h}_{\mathsf{cmt}}$*, he may point out that B hasn't answered his question or acknowledge B's response having been assured of a win.*
- $(\mathsf{tp}_3, \sigma_{\mathsf{qelab}})$*: Here P believes that the Jury expects him to follow up on his initial question and he obliges with* σ_{qelab}*.*

In the actual conversation, since P did acknowledge B's response π_5*, we conclude that P was either type* tp_1 *and believed that the Jury was of type* tj_1 *or he was of type* tp_2 *and believed that the Jury was of type* tj_2*. Hence,* given his beliefs*, P was rational in acknowledging B's response.*

By assigning the Jury types, we have explained the conversational behavior of Bronston and the Prosecutor in a much more rigorous and satisfactory way than is possible in [3] or [5]. There is an additional advantage to assigning the Jury types. We can now introduce uncertainty on the part of the players as to the Jury's type. This means we can model the fact that players cannot reliably estimate neither the Jury winning conditions, nor the number of turns after which the Jury will decide to end the game and not even the weighting function of the Jury (if it uses one for the estimation of the winning conditions as described in [4]). They can only form beliefs about these parameters and strategize accordingly. This prevents troublesome Backwards Induction arguments discussed in [5], in cases where it is common knowledge that the ME game is finite.

4.5 A more complex example

We have now introduced our key epistemic concepts and illustrated them with our simple Bronston example. In that example, the updates of beliefs are rather trivial. We now turn to a more complex example, also discussed in [5], to illustrate the interdependence of a player's interpretation function and belief function.

Example 8 *As background to this excerpt from a press conference by Senator Coleman's spokesman Sheehan, Senator Coleman was running for re-election as a senator from Minnesota in the 2008 US elections.*

S.a **Reporter:** *On a different subject is there a reason that the Senator won't say whether or not someone else bought some suits for him?*
S.b **Sheehan:** *Rachel, the Senator has reported every gift he has ever received.*
S.c **Reporter:** *That wasn't my question, Cullen.*
S.d **Sheehan:** *(i) The Senator has reported every gift he has ever received. (ii) We are not going to respond to unnamed sources on a blog.*
S.e **Reporter:** *So Senator Coleman's friend has not bought these suits for him? Is that correct?*
S.f **Sheehan:** *The Senator has reported every gift he has ever received.*

Sheehan continues to repeat, *The Senator has reported every gift he has ever received* seven more times in two minutes to every follow up question by the reporter corps. `http://www.youtube.com/watch?v=VySnpLoaUrI`

To formulate (8) as an ME game, we assume two active players (i) the reporter corps (R) and (i) spokesman Sheehan (S). The play in (8), as well as some alternative moves, exploits the following EDU characterizations.

1. **R:** $\langle \pi_0$: *On a different subject is there a reason that the Senator won't say whether or not someone else bought some suits for him?* $(\phi_0)\rangle$
2. **R:** $\langle \pi_2$: *That wasn't my question, Cullen.* $(\phi_2)\rangle$
3. **R:** $\langle \pi_5$: *So Senator Coleman's friend has not bought these suits for him?* $(\phi_5)\rangle$ $\langle \pi_6$: *Is that correct?* $(\phi_6)\rangle$
4. **R:** $\langle \pi_2^2$:= *Ok.* $(\phi_2^2)\rangle$
5. **S:** $\langle \pi_\alpha^1$: *Rachel, the Senator has reported every gift he has ever received.* $(\phi_\alpha)\rangle$
6. **S:** $\langle \pi_\alpha^2$: *The Senator has reported every gift he has ever received.* $(\phi_\alpha)\rangle$ $\langle \pi_3$: *We are not going to respond to unnamed sources on a blog.* $(\phi_3)\rangle$
7. **S:** $\langle \pi_\alpha^3$:*The Senator has reported every gift he has ever received.* $(\phi_\alpha)\rangle$
8. **S:** $\langle \pi_1^2$:=*Yes.* $(\phi_1^2)\rangle$
9. **S:** $\langle \pi_1^3$:=*No.* $(\phi_1^3)\rangle$

The ULF given below for (8) represents a play in an ME game, which we denote as ρ.

$$\begin{aligned}\rho =&\langle \pi_0 : \phi_0\rangle\langle \pi_1 : \pi_0\pi_\alpha^1 \exists \mathcal{R}_1.\mathcal{R}_1(\pi_0,\pi_\alpha^1)\rangle\langle \pi_2 : \pi_2^1 \mathsf{corr}(\pi_1,\pi_2^1)\rangle\\ &\langle \pi_8 : \langle \pi_4 : \pi_\alpha^2\pi_3 \mathsf{exp}(\pi_\alpha^2,\pi_3)\exists \mathcal{R}_2 \exists x.\mathcal{R}_2(x,\pi_4)\rangle\rangle\\ &\langle \pi_9 : \langle \pi_7 : \pi_5\pi_6 \mathsf{confQ}(\pi_5,\pi_6)\rangle \exists z.\mathsf{res}(z,\pi_7)\rangle\langle \pi_{10} : \pi_\alpha^3 \exists \mathcal{R}_3 \exists y.\mathcal{R}_3(y,\pi_\alpha^3)\rangle\end{aligned}$$

$|\rho| = 6$ and let $\rho_0, \rho_1, \ldots, \rho_6$, where $\rho_0 = \epsilon$ be the prefixes of ρ after $0, 1, \ldots, 6$ turns of the game respectively (that is, $\rho_6 = \rho$). Note that the ULF ρ has three underspecified relations $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ as well as two underspecified arguments. The CDU π_1 groups together π_0 and π_α^1 as well as the underspecified relation $\mathcal{R}_1$, which is clearly the target of the correction prompted by π_2^1. In SDRT, to target a relation instance that is corrected, the correction must take scope over a

CDU containing that relation instance. It is also uncontroversial that explanation holds between π^2_α and π_3 and that a confirmation question relation holds between π_5 and π_6. On the other hand it is unclear how to connect either π^2_α or the CDU π_4 to the prior discourse structure; the same situation holds for π^3_α. We have noted that the left arguments of these two underspecified relations is itself underspecified by existentially quantifying over those arguments. The explicit discourse connector *So* signals a result between π_4, π_8, or some other DU and the CDU π_7, formed from π_5 and π_6.

ρ serves as the spine of the game tree depicted in Figure 3. The relation instances whose arguments are underspecified are drawn in red.

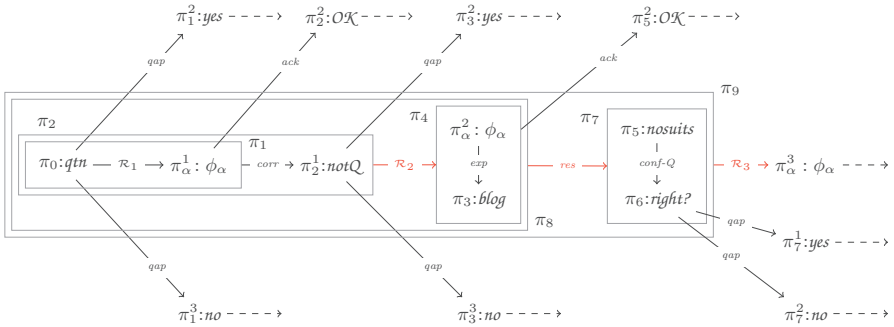

Fig. 3: An ME game structure for the initial part of the conversation between R and S.

Now, we can imagine two types for Sheehan: t_H and t_D.

- t_H is the 'honest' type, according to which Sheehan truly implicates that the Senator did not receive the suits and that he simply does not want to respond to this charge.
- t_D is the 'dishonest' type, according to which Coleman received the suits but did not declare them and Sheehan is trying to cover this fact up.

We can identify the Jury with the reporter corps or in general with the audience present in that room during the interview or even with everyone for whom the interview is relevant. We envisage two types for the Jury: tj_U and tj_B.

- tj_U is the 'unbiased' Jury that starts out with a presumption of full disclosure and honesty from Sheehan.
- tj_B is the type of the Jury that is disposed to believe whatever the spokesman says, because, for example, they are from the same political party. Such a Jury is 'biased'.

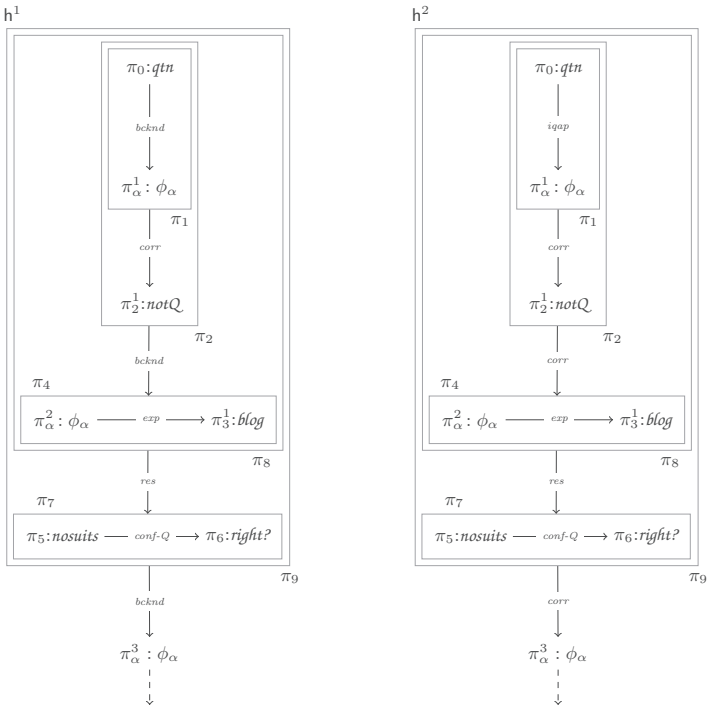

Fig. 4: Two histories resulting from two different interpretations of the uninterpreted relations in ρ

Now, ρ can have at least two different interpretations giving rise to two different histories h^1 and h^2.

$$\begin{aligned}\mathsf{h}^1 =&\langle\pi_0 : \phi_0\rangle\langle\pi_1 : \pi_0\pi^1_\alpha\mathsf{bcknd}(\pi_0,\pi^1_\alpha)\rangle\langle\pi_2 \colon \pi^1_2\mathsf{corr}(\pi_1,\pi^1_2)\rangle\\ &\langle\pi_8 \colon \langle\pi_4 \colon \pi^2_\alpha\pi_3\mathsf{exp}(\pi^2_\alpha,\pi_3)\mathsf{bcknd}(\pi_2,\pi_4)\rangle\rangle\\ &\langle\pi_9 \colon \langle\pi_7 : \pi_5\pi_6\mathsf{confQ}(\pi_5,\pi_6)\rangle\mathsf{res}(\pi_8,\pi_7)\rangle\langle\pi_{10} \colon \pi^3_\alpha\mathsf{bcknd}(\pi_9,\pi^3_\alpha)\rangle\end{aligned}$$

$$\begin{aligned}\mathsf{h}^2 =&\langle\pi_0 : \phi_0\rangle\langle\pi_1 : \pi_0\pi^1_\alpha\mathsf{iqap}(\pi_0,\pi^1_\alpha)\rangle\langle\pi_2 \colon \pi^1_2\mathsf{corr}(\pi_1,\pi^1_2)\rangle\\ &\langle\pi_8 \colon \langle\pi_4 \colon \pi^2_\alpha\pi_3\mathsf{exp}(\pi^2_\alpha,\pi_3)\mathsf{corr}(\pi_2,\pi_4)\rangle\rangle\\ &\langle\pi_9 \colon \langle\pi_7 : \pi_5\pi_6\mathsf{confQ}(\pi_5,\pi_6)\rangle\mathsf{res}(\pi_8,\pi_7)\rangle\langle\pi_{10} \colon \pi^3_\alpha\mathsf{corr}(\pi_9,\pi^3_\alpha)\rangle\end{aligned}$$

We can depict these histories graphically as in Figure 4. We let $\mathsf{h}^1_j, \mathsf{h}^2_j,\ j \in \{0,1,\ldots,6\}$ be the prefixes of h^1 and h^2 respectively after j turns where $\mathsf{h}^1_0 = \mathsf{h}^2_0 = \epsilon$.

The types tj_U and tj_B of the Jury have different priors concerning S's types. tj_U starts with an indifference between t_H and t_D, while tj_B starts off believing S is of type t_H, with a high probability (say 0.7). We assume that the beliefs of

the Jury are independent of the strategies of S. Then their beliefs on the types of S can be given as in Table 4a. We also assume that throughout the course of the game, whenever a Jury interprets ρ or any prefix of it as history h^1 it is indifferent between the types t_H and t_D of S and whenever it interprets ρ or any prefix of it as history h^2 it assigns a higher probability to the type t_H of S. This is irrespective of the type of the Jury itself. This is represented in Table 4a where $j \in \{0, 1, \ldots, 6\}$.

$\beta_{\mathcal{J}}^{\rho_0}(\ \cdot\ , \epsilon)$	t_H	t_D
tj_U	0.5	0.5
tj_B	0.7	0.3

$\beta_{\mathcal{J}}^{\rho_j}(\cdot)$	t_H	t_D
h^1	0.5	0.5
h^2	0.7	0.3

Table 4a: The beliefs of the Jury about the type of S before the start of the game and the beliefs of the Jury about the type of S throughout the play ρ

Now as the play ρ progresses, the types tj_U and tj_B of the Jury end up interpreting ρ differently, as h^1 and h^2 respectively. These interpretations are intuitively justified as follows. We will give a qualitative analysis for readability, though we could have given actual numbers for the probability values and computed the belief updates in every step.

– **Beliefs and interpretations of type tj_U:** When type tj_U updates with the unexpected ϕ_α as a response to $\langle \pi_0 : \mathit{qtn} \rangle$, it is genuinely puzzled by the response. While it is natural to assume that an honest senator has never received any gifts from a friend which he has not reported, the inference from ϕ_α as an answer to π_0, as to why the Senator has not said anything about the suits, is complicated and indirect. A Jury must consider the interpretation of S.a and S.b conditioned on both t_D and t_H. Conditioning on the assumption that S is of type t_D and S's response α, the Jury, like R, assigns a much higher probability to the interpretation illustrated in h^1, that S.b does not answer S.a and is rather related to it via background. That is,

$$\xi_{\mathcal{J}}^{\rho_2}(\mathsf{tj}_U, t_D)(\mathsf{h}_2^1) \gg \xi_{\mathcal{J}}^{\rho_2}(\mathsf{tj}_U, t_D)(\mathsf{h}_2^2)$$

On the other hand, conditioning on the assumption that S is of type t_H and the response α, the Jury confers only a slightly higher probability to an iqap relation than a background relation between S.a and S.b. That is,

$$\xi_{\mathcal{J}}^{\rho_2}(\mathsf{tj}_U, t_H)(\mathsf{h}_2^1) > \xi_{\mathcal{J}}^{\rho_2}(\mathsf{tj}_U, t_H)(\mathsf{h}_2^2)$$

When we combine the probabilities over t_D and t_H—because tj_1 is considering both—we get a higher probability for background than for iqap, leading to a higher probability of h_2^1.

$$\xi_{\mathcal{J}}^{\rho_2}(\mathsf{tj}_U)(\mathsf{h}_2^1) > \xi_{\mathcal{J}}^{\rho_2}(\mathsf{tj}_U)(\mathsf{h}_2^2)$$

Combining this with how the believes are derived as given by Table 4a, we can conclude that:

$$\beta^{\rho_2}_{\mathcal{J}}(\mathsf{tj}_U)(t_H) > \beta^{\rho_2}_{\mathcal{J}}(\mathsf{tj}_U)(t_D)$$

Next, conditioning in turn on this belief, tj_U naturally interprets R's response π^1_2 as a correction of S's move as implicating any kind of answer and hence implicating R's request for a direct answer to π_0. In π^2_α, however, S reiterates his original response, and explains why he does so in π_3: the Senator and his staff do not want to comment on unnamed sources on some blog. So at this point tj_U might lean back towards h^2_4, interpreting π_4 as correcting the exchange in π_2. Hence we have $\xi^{\rho_4}_{\mathcal{J}}(\mathsf{tj}_U)(\mathsf{h}^2_4) = 1$. tj_U then takes up the natural conclusion from ϕ_α as an iqap to π_0, which would be the upshot of S's correction of R's correction—namely, that S had in fact replied to R's question in π^1_α. This is shown in both h^1 and h^2 by linking π_8 to π_7 and marking the relation between them as result. R also follows up with a confirmation question to S that this is so (π_6). At this point we still have $\xi^{\rho_5}_{\mathcal{J}}(\mathsf{tj}_U)(\mathsf{h}^2_5) = 1$. However to this, S replies with ϕ_α once again in (S.f), which yields the EDU π^3_α. Now tj_U is confused. Why is S not replying with a direct answer *yes* or *no*? Is the Senator in fact dishonest, of type t_D, and S is trying to hide this fact? tj_U shift backs to the history h^1, and treat the links between π_4 and π_2 and between π^3_α and π_9 as background. The belief of tj_U about the type of S has shifted towards t_D now, which means ϕ_α is taken as an evading of the question. We would thus have $\xi^{\rho}_{\mathcal{J}}(\mathsf{h}^1) = 1$ and this, according to Table 4a, leads to $\beta^{\rho}_{\mathcal{J}}(\mathsf{tj}_U)(t_H) = \beta^{\rho}_{\mathcal{J}}(\mathsf{tj}_U)(t_D) = 0.5$.

– **Beliefs and interpretations of type tj_B:** Given its high confidence that S is of type t_H, the Jury type tj_B accepts ϕ_α as a perfectly acceptable indirect answer to π_0 and so opts for the history h^2's interpretation of that first underspecified relation.

$$\xi^{\rho_2}_{\mathcal{J}}(\mathsf{tj}_B, t_H)(\mathsf{h}^2_2) \gg \xi^{\rho_2}_{\mathcal{J}}(\mathsf{tj}_B, t_H)(\mathsf{h}^1_2)$$

It would also interpret the relation $\mathcal{R}_2$ of ρ as a correction as in h^2, it would construct a different history after π_6. It would see each repetition of ϕ_α as another correction of R's attempts to reopen a topic that that S has already settled. Since S is of type t_H (with a high probability), he need not continue the discussion of a matter that has already been labeled as one that Sheehan will not comment on. Thus after ρ, $\xi^{\rho}_{\mathcal{J}}(\mathsf{tj}_B)(\mathsf{h}^2) = 1$ and hence by Table 4a, $\beta^{\rho}_{\mathcal{J}}(\mathsf{tj}_B)(t_H) = 0.7$ and $\beta^{\rho}_{\mathcal{J}}(\mathsf{tj}_B)(t_D) = 0.3$.

Next, let us analyse the conversation as it proceeded after (S.e). S in effect refuses to engage with R by repeating ϕ_α to every follow up question on the topic. We see how this explicit linguistic uncooperativity (in the sense of [3]) affects the Jury's estimation of the Senator's type, given that it keeps revising its beliefs according to Bayesian updates.

Let us assume that every relevant ensuing coherent move by R consists of the single question EDU:

ϕ_Q =*Has the Senator received gifts from his friend?*

This is a simplification of what actually happened but all the actual questions were in fact follow up questions to ϕ_Q or questions related to it. So to simplify the presentation, we'll treat them all as question ϕ_Q.

S has three consistent coherent moves: ϕ_1^2, ϕ_1^3 and ϕ_α where

- ϕ_1^2 is a positive response from S: *yes, the Senator has received gifts from his friend.*
- ϕ_1^3 is a negative response: *no, the Senator has never received gifts from his friend.*
- ϕ_α is the response: *the Senator has reported every gift he has ever received.*

The ME game after (S.f) looks as shown in Figure 5 where again the dashed edges represent continuations which are irrelevant to the present analysis.

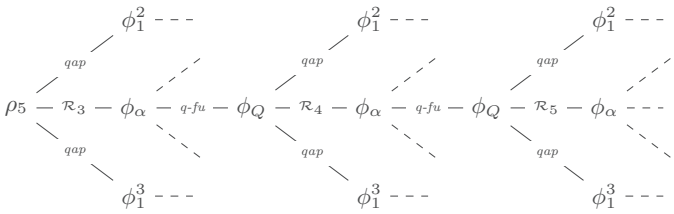

Fig. 5: The ME game of Eg. 8 after ρ_5

Although S repeats ϕ_α 7 more times after ρ_5 in the press conference containing (8), for simplicity of this analysis, we shall consider only 3 rounds after play of the above game. Let ρ^6, ρ^7 and ρ^8 be the extension of ρ_5 after each of these rounds, where $\rho^6 = \rho$. As before, each of these plays can be interpreted in two different ways: (i) histories of the form h_j^1 where the relations $\mathcal{R}_3, \mathcal{R}_4$ and $\mathcal{R}_5$ are interpreted as bcknd and (ii) histories of the form h_j^2 where the relations $\mathcal{R}_3$, $\mathcal{R}_4$ and $\mathcal{R}_5$ are interpreted as corr.

There are 7 strategies of S that are relevant for these three rounds which are given by the set: $S_S = \{\sigma_1, \sigma_2, \ldots, \sigma_7\}$ and are presented in Table 4b.

	round 1	round 2	round 3
σ_1	$\phi_1^2 \equiv$*yes*	–	–
σ_2	$\phi_1^3 \equiv$*no*	–	–
σ_3	ϕ_α	$\phi_1^2 \equiv$*yes*	–
σ_4	ϕ_α	$\phi_1^3 \equiv$*no*	–
σ_5	ϕ_α	ϕ_α	$\phi_1^2 \equiv$*yes*
σ_6	ϕ_α	ϕ_α	$\phi_1^3 \equiv$*no*
σ_7	ϕ_α	ϕ_α	ϕ_α

Table 4b: The relevant strategies of S after the play ρ

Let us now look at how each of the types tj_U and tj_B of the Jury would update its beliefs about S's type given the course of the conversation after ρ_5.

Jury type $\mathbf{tj}_U$ We saw that tj_U is the fair type that ends up interpreting ρ as h^1 with turn (S.f). It starts off the game believing with a probability of 0.5 that S is of an honest type t_H, which it maintains after the two of responses ϕ_α by S in ρ. That is, it assigns an equal probability to S being of type t_H or t_D after (S.f). We assume that such tj_U sticks to this interpretation also after the rounds following ρ. That is, it interprets ρ^7 and ρ^8 as h^1_7 and h^1_8. However, such a Jury type would expect that if S is indeed of type t_H then he would eventually give the direct answer $\phi^7_1 \equiv$*no* to the confirmation question in π_6. In addition, for simplicity, suppose that tj_U believes that is it equally likely for S to give a direct answer to ϕ_Q in any of the three rounds that we have considered after (S.f). Given all the above assumptions, we represent the beliefs of the Jury type tj_U after the play ρ as shown in Table 4c(i) (with all calculations to 3-decimal precision).

$\beta^\rho_{\mathcal{J}}(\mathsf{tj}_U)$	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7
t_H	0	0.167	0	0.167	0	0.166	0
t_D	0.125	0	0.125	0	0.125	0	0.125

Table 4c(i): The beliefs of the Jury type tj_U about the type-strategy pair of S after ρ

Now, let E^ρ_H be the $\mathcal{J}$-event that S is of type t_H and E^ρ_D be the $\mathcal{J}$-event that he is of type t_D. Formally, $E^\rho_H = \{t_H\} \times S_S$ and $E^\rho_D = \{t_D\} \times S_S$. After the play ρ we have that $\beta^\rho_{\mathcal{J}}(\mathsf{tj}_U)(E^\rho_H) = \beta^\rho_{\mathcal{J}}(\mathsf{tj}_U)(E^\rho_D) = 0.5$.

The strategies of S that are compatible with the play ρ^7 are $S^7_S = \{\sigma_3, \sigma_4, \sigma_5, \sigma_6, \sigma_7\}$. Hence, we can define the $\mathcal{J}$-events $E^{\rho^7}_H = \{t_H\} \times S^7_S, E^{\rho^7}_D = \{t_D\} \times S^7_S$ and $E^{\rho^7} = E^{\rho^7}_H \cup E^{\rho^7}_D$.

Now, $\beta^\rho_{\mathcal{J}}(\mathsf{tj}_U)(E^{\rho^7}) = 0.708$. Suppose the Jury derives its beliefs after ρ^7 by performing a Bayesian update of its beliefs after ρ. Let $j \in \{4, 6\}$. Then we have

$$\beta^{\rho^7}_{\mathcal{J}}(\mathsf{tj}_U)(\langle t_H, \sigma_j\rangle) = \beta^\rho_{\mathcal{J}}(\mathsf{tj}_U)(\langle t_H, \sigma_j\rangle \mid E^{\rho^7}) = 0.167/0.708 = 0.238$$

and for $k \in \{3, 5, 7\}$

$$\beta^{\rho^7}_{\mathcal{J}}(\mathsf{tj}_U)(\langle t_D, \sigma_k\rangle) = \beta^\rho_{\mathcal{J}}(\mathsf{tj}_U)(\langle t_D, \sigma_k\rangle \mid E^{\rho^7}) = 0.125/0.708 = 0.175$$

Thus after the first round of the repetition of ϕ_α by S, the beliefs of Jury type tj_U, after Bayesian updates, can be represented as shown in Table 4c(ii).

We have $\beta^{\rho^7}_{\mathcal{J}}(\mathsf{tj}_U)(E^{\rho^7}_H) = 0.476$ and $\beta^{\rho 7}_{\mathcal{J}}(\mathsf{tj}_U)(E^{\rho^7}_D) = 0.525$ (as can be seen by summing the individual rows of Table 4c(ii)).

$\beta_{\mathcal{J}}^{\rho^7}(\mathsf{tj}_U)$	σ_3	σ_4	σ_5	σ_6	σ_7
t_H	0	0.238	0	0.238	0
t_D	0.175	0	0.175	0	0.175

Table 4c(ii): The beliefs of the Jury type tj_U after ρ^7

Next, the strategies that are compatible with ρ^8 are $S_S^8 = \{\sigma_5, \sigma_6, \sigma_7\}$. As before, we can define the events $E_H^{\rho^8} = \{t_H\} \times S_S^8, E_D^{\rho^8} = \{t_D\} \times S_S^8$ and $E^{\rho^8} = E_H^{\rho^8} \cup E_D^{\rho^8}$ and hence $\beta_{\mathcal{J}}^{\rho 7}(\mathsf{tj}_U)(E^{\rho^8}) = 0.587$. We have, as before

$$\beta_{\mathcal{J}}^{\rho^8}(\mathsf{tj}_U)(\langle t_H, \sigma_6\rangle) = \beta_{\mathcal{J}}^{\rho^7}(\mathsf{tj}_U)(\langle t_H, \sigma_6\rangle \mid E^{\rho^8}) = 0.238/0.587 = 0.404$$

and for $j \in \{5, 7\}$

$$\beta_{\mathcal{J}}^{\rho^8}(\mathsf{tj}_U)(\langle t_D, \sigma_k\rangle) = \beta_{\mathcal{J}}^{\rho^7}(\mathsf{tj}_U)(\langle t_D, \sigma_j\rangle \mid E^{\rho^8}) = 0.175/0.587 = 0.298$$

Thus, $\beta_{\mathcal{J}}^{\rho^8}(\mathsf{tj}_U)(E_H^{\rho^8}) = 0.404$ and $\beta_{\mathcal{J}}^{\rho^8}(\mathsf{tj}_U)(E_D(\rho^8)) = 0.596$. So after round 2, and after Bayesian updates, the type tj_U of the Jury believes even more that S is of type t_D and not of type t_H.

The beliefs of tj_U after each round of the conversation can be represented pictorially as shown in Figure 6.

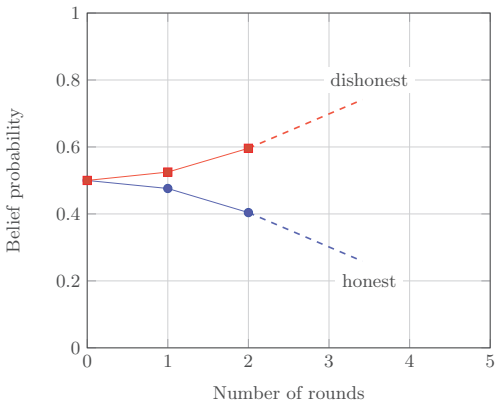

Fig. 6: The progressive change in the beliefs of the Jury type tj_U about the type of S

Given these calculations, we can imagine that a Jury of type tj_U might then stop the conversation once the probability of t_D becomes high enough. For such a Jury, S's repetitions doom his play to be losing.

Jury type tj_B We next perform a similar analysis for the Jury type tj_B which has a different interpretation of S's repeated responses ϕ_α. As we saw in our earlier discussion, such a Jury type interprets the play ρ as the history h^2. It is ready to believe that S has indeed "settled the topic" with his response π_3. Any attempt to re-open the topic would simply be unneccessary and hence such a Jury type is perfectly happy with S's response ϕ_α to R's repeated question ϕ_Q. We argued that after the play ρ the Jury type tj_B comes away assigning a rather high probability (0.7 say) to S being of the honest type t_H. As in the previous case we again analyse the game for three more rounds after ρ again assuming that S has the strategy set S_S at his disposal. We assume that such tj_B sticks to its interpretation also after the rounds following ρ and interprets ρ^7 and ρ^8 as h^2_7 and h^2_8. tj_B takes S's response ϕ_α as being compatible with his type t_H. The only case where S would reveal his type to be t_D is when he gives the direct answer $\phi^2_1 \equiv$*yes* to R's question ϕ_Q. Hence, this time, we can represent the beliefs of the Jury type tj_B after the play ρ as shown in Table 4d(i).

$\beta[\rho]_{\mathcal{J}}(\mathsf{tj}_B)$	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7
t_H	0	0.175	0	0.175	0	0.175	0.175
t_D	0.1	0	0.1	0	0.1	0	0

Table 4d(i): The beliefs of the Jury type tj_U about the type-strategy pair of S after ρ given the history h^1

Let, as in the previous case, E^ρ_H be the event that S is of type t_H and $E^{\mathcal{J}}_D$ be the event that he is of type t_D. That is, $E^\rho_H = \{t_H\} \times S_S$ and $E_D = \{t_D\} \times S_S$. After the play ρ we have that $\beta^\rho_{\mathcal{J}}(\mathsf{tj}_B)(E^\rho_H) = 0.7$ and $\beta^\rho_{\mathcal{J}}(\mathsf{tj}_B)(E^\rho_D) = 0.3$.

Once again, the strategies of S that are compatible with ρ^7 are $S^7_S = \{\sigma_3, \sigma_4, \sigma_5, \sigma_6, \sigma_7\}$. We define the events $E^{\rho^7}_H = \{t_H\} \times S^7_S, E^{\rho^7}_D = \{t_D\} \times S^7_S$ and $E^{\rho^7} = E^{\rho^7}_H \cup E^{\rho^7}_D$. Now, $\beta^\rho_{\mathcal{J}}(\mathsf{tj}_B)(E^{\rho^7}) = 0.725$. Let $j \in \{4, 6, 7\}$. Then we have

$$\beta^{\rho^7}_{\mathcal{J}}(\mathsf{tj}_B)(\langle t_H, \sigma_j\rangle) = \beta^\rho_{\mathcal{J}}(\mathsf{tj}_B)(\langle t_H, \sigma_j\rangle \mid E^{\rho^7}) = 0.175/0.725 = 0.241$$

and for $k \in \{3, 5\}$

$$\beta^{\rho^7}_{\mathcal{J}}(\mathsf{tj}_B)(\langle t_D, \sigma_k\rangle) = \beta^\rho_{\mathcal{J}}(\mathsf{tj}_B)(\langle t_D, \sigma_k\rangle \mid E^{\rho^7}) = 0.1/0.725 = 0.138$$

Thus after round 1, the beliefs of Jury type tj_B, after Bayesian updates, can be represented as shown in Table 4d(ii).

We have $\beta^{\rho^7}_{\mathcal{J}}(\mathsf{tj}_B)(E^{\rho^7}_H) = 0.724$ and $\beta^{\rho^7}_{\mathcal{J}}(\mathsf{tj}_B)(E^{\rho^7}_D) = 0.276$ (as can be seen by summing the individual rows of Table 4d(ii)).

Next, the strategies that are compatible with ρ^8 are $S^8_S = \{\sigma_5, \sigma_6, \sigma_7\}$. As before, we can define the events $E^{\rho^8}_H = \{t_H\} \times S^8_S, E^{\rho^8}_D = \{t_D\} \times S^8_S$ and $E^{\rho^8} =$

$\beta_{\mathcal{J}}^{\rho^7}(\mathsf{tj}_B)$	σ_3	σ_4	σ_5	σ_6	σ_7
t_H	0	0.241	0	0.241	0.242
t_D	0.138	0	0.138	0	0

Table 4d(ii): The beliefs of the Jury type tj_B after ρ^7

$E_H^{\rho^8} \cup E_D^{\rho^8}$ and hence $\beta_{\mathcal{J}}^{\rho^7}(\mathsf{tj}_B)(E^{\rho^8}) = 0.620$. We have, as before, for $j \in \{6, 7\}$

$$\beta_{\mathcal{J}}^{\rho^8}(\mathsf{tj}_B)(\langle t_H, \sigma_j\rangle) = \beta_{\mathcal{J}}^{\rho^7}(\mathsf{tj}_B)(\langle t_H, \sigma_j\rangle \mid E^{\rho^8}) = 0.241/0.620 = 0.389$$

and

$$\beta_{\mathcal{J}}^{\rho^8}(\mathsf{tj}_B)(\langle t_D, \sigma_5\rangle) = \beta^{\rho^7}(\mathsf{tj}_B)(\langle t_D, \sigma_5\rangle \mid E^{\rho^7}) = 0.138/0.620 = 0.222$$

Thus, $\beta_{\mathcal{J}}^{\rho^8}(\mathsf{tj}_B)(E_H^{\rho^8}) = 0.778$ and $\beta_{\mathcal{J}}^{\rho^8}(\mathsf{tj}_B)(E_D^{\rho^8}) = 0.2226$. So after round 2, and after the Bayesian updates, the belief of the type tj_B that S is of the honest type t_H is strengthened even further.

The beliefs of tj_B after each round of the conversation can be represented pictorially as shown in Figure 7.

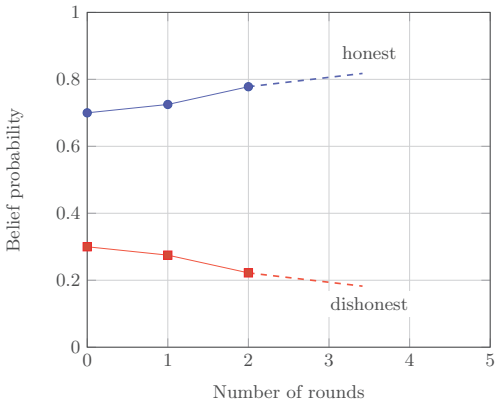

Fig. 7: The progressive change in the beliefs of the Jury type tj_B about the type of S

4.6 Confirmation of bias

Example (8) gives us an insight into the nature of the evaluations of strategic conversations by different types of Juries. While the Jury of type tj_F was arguably guided by facts and evidence in deriving its beliefs and updating them,

tj_B was more inclined to favour S in its evaluation of the exchange. It had preconceived beliefs about the type of S being honest and interpreted his messages ϕ_α as such. These interpretations, in turn, served to further strengthen its beliefs about the type of S being tj_B which then influenced its interpretations of ϕ_α more and more in the later stages of the exchange.

This points to an interesting and intuitive truth about interpretation: a Jury with a prior disposition towards a player type is guided by it in its interpretation of the messages in the conversation. Surprisingly, this in turn serves to strengthen its disposition further in turn effecting its subsequent interpretations of the conversation. We formalise this phenomenon as follows.

Suppose there are two types for a player t_1 and t_2 and let ρ be a play of the ME game. Suppose ρ can be interpreted in two strands of histories h^1 and h^2. Suppose $t_{\mathcal{J}}$ is a Jury type whose beliefs about the type of the player are given as in Table 5. $\epsilon : 0 < \epsilon \leq 1/4$ is called the index of bias. The table says that as the Jury type $t_{\mathcal{J}}$ interprets longer and longer prefixes of ρ as h^1, it is more and more likely to believe that the player is of type t_1.

$\beta_{\mathcal{J}}^{\rho_j}(t_{\mathcal{J}},\ \cdot\ )$	t_1	t_2
h_j^1	$0.5 + \sum_{k=1}^{j} \epsilon^j$	$0.5 - \sum_{k=1}^{j} \epsilon^j$
h_j^2	0.5	0.5

$\xi_{\mathcal{J}}^{\rho_j}(t_{\mathcal{J}},\ \cdot\ )$	h_j^1	h_j^2
t_1	1	0
t_2	0	1

Table 5: Beliefs and interpretations of the Jury type $t_{\mathcal{J}}$

Also, suppose t_j interprets the prefixes of ρ as given in Table 5. That is, if $t_{\mathcal{J}}$ believes that the player is of type t_1 then it interprets the current prefix of ρ as h^1 and if it believes that the player type is t_2 then it interprets it as h^2.

Now, suppose $\epsilon = 0.25$ and suppose $t_{\mathcal{J}}$ interprets the play ρ_1 (after the first turn) as h_1^1. From Table 5 we have that $t_{\mathcal{J}}$ believes with probability $0.5+0.25^1 = 0.75$ that the player is of type t_1 and with a probability 0.25 that she is of type t_2. This means that after the next turn, from Table 5 we have that it is likely to interpret ρ_2 as h_2^1 with the higher probability of 0.75 and as h_2^2 with the lower probability of 0.25. So suppose $t_{\mathcal{J}}$ does interpret ρ_2 as h_2^1. Then again from Table 5 we have that $t_{\mathcal{J}}$ believes with probability $0.5 + 0.25^1 + 0.25^2 = 0.875$ that the player is of type t_1 and with a probability 0.125 that she is of type t_2. This means that after the next turn, from Table 5 we have that it is likely to interpret ρ_3 as h_3^1 with the higher probability of 0.875 and as h_3^2 with the lower probability of 0.125. We can carry out a similar analysis for all the ensuing turns.

This leads us to make the following general observation.

Observation 1 (Confirmation of bias) *In a strategic conversation if a Jury assigns a higher belief probability to a particular type t of a player (participant) and if it interprets the underspecified discourse relations of the play of the ME game corresponding to the conversation in tune with its beliefs about the player types then such interpretations strengthen its belief on t. This in turn shifts its*

beliefs further towards t in the consecutive rounds of the conversation biasing its interpretations of the underspecified relations in the subsequent rounds even further.

In our treatment of Example 8, we saw already a case of a Jury, the one of type tj_B, with a predisposition to a particular type for one of the players. It is a special case of our generalization above. However, we also saw that a Jury without such a predisposition, the one of type tj_U, could also fix upon a certain history and then instill and reinforce a particular view of one of the players. As we argued, tj_U also changes its beliefs as to which history is more probable during the course of interpretation. This happened because tj_U took the linguistic cues, in particular the repetition of α in response to its attempt to confirm h^2, to shift its belief in h^2 to a belief in h^1, which in turn reinforced the probability of Sheehan as t_D. It is possible, however, for a Jury to switch from a history h to h' even if the prior probability of h is high, provided the linguistic evidence is sufficiently strong in favor of h'.

A very important question at this point comes to mind:

Question 1 *When is a Jury* biased*? What constitutes an* unbiased Jury*?*

We do not have a definite answer to what an unbiased Jury is. An unbiased Jury gives all (relevant) player types an equal chance and interpret plays in light of this. Secondly, an unbiased Jury is indifferent about the identities of the players. That is to say that if it assigns win to Player i for a play ρ, then it should assign win to Player $(1-i)$ for the play which has the exact same discourse moves and relations as ρ but where the roles of the players have been interchanged. We thus list here two necessary conditions that an unbiased Jury must intuitively satisfy.

Towards that we first define the notion of the dual of a play of an ME game. Let $(v, i) \in (V_0 \cup V_1)^\omega$ be an element of the labeled vocabulary. Define its dual as:

$$\overline{(v, i)} = (v, 1-i)$$

The dual of a play $\rho \in ((V_0 \cup V_1)^\omega)^\infty$ then is simply the lifting of this operator over the entire sequence of ρ. That is, if $\rho = x_0 x_1 x_2 \ldots$, where $x_0 = \epsilon$ then

$$\overline{\rho} = x_0 \overline{x_1}\ \overline{x_2} \ldots$$

We now state the two constraints that an unbiased Jury must necessarily satisfy.

- **Indifference towards player identity:** A Jury $\mathcal{J} = (Win_0, Win_1)$ is unbiased if for every $\rho \in (V_0 \cup V_1)^\omega$, $\rho \in Win_i$ iff $\overline{\rho} \in Win_{(1-i)}$.
- **Symmetry of prior belief:** A Jury is unbiased if it has symmetrical prior beliefs about the player types and no bias in its interpretation of the plays of the ME game. For instance, in our above analysis, such a Jury would be characterized as having $p = 0.5$ in its prior beliefs about the player types and $\epsilon = 0$.

An unbiased Jury is thus indifferent between the identity of the participants and evaluates a conversation based solely on the strength of the points put forth by them. A biased Jury however, has a more 'selective' listening. It is often blind to the inconsistencies and the factual errors of the participants towards whom it is biased. There are numerous examples where the 2016 Presidential candidate Trump asserted inconsistencies and factual errors throughout his campaign. This did not stop his supporters from voting for him in anyway.

The second necessary condition above for an unbiased Jury leads to a remarkable conclusion. Assuming that the prior beliefs of the Jury are assigned uniformly at random (or according to some other continuous distribution), a simple measure-theoretic argument convinces us that

Observation 2 (Biased Jury) *Almost surely, a Jury is always biased.*

Signaling games in retrospect We end this section with a brief comparison of our analysis of the subjectivity of interpretation and that given by signaling games. signaling games typically analyze the meaning of a signal in terms of reflective equilibrium. As this equilibrium depends on the beliefs of the players, signaling games also predict that interpretation can be subjective. Epistemic ME games take the meaning, which may include implicatures, of elements of a play to be exogenously determined; ME games exploit linguistic theory to constrain what elements in the signal are subject to linguistic interpretation. As we mentioned earlier, the assumption of exogenous meaning is necessary to get a well-defined interpretation in cases where players' interests are opposed, as argued in detail in [5].

The use of SDRT in ME games, a formal theory of discourse interpretation, isolates a crucial component for subjective and biased interpretation that is missing in standard signaling games: the way an interpreter links the discourse units in a given play in terms of discourse relations. Such links are crucial to the Jury's bias in the analysis of example (8). This point generalizes to other, lexical ambiguities. But as discourse connections ultimately determine whether the speaker is making a relevant contribution to the conversation and are often underspecified, they seem particularly apt at generating biased interpretations.

A final difference between signaling games and epistemic ME games is that signaling games have a built in asymmetry in their treatment of the receiver and sender. The sender has complete information but the receiver does not. In an epistemic ME game, this asymmetry is no longer present, as intuitions would dictate; as conversation proceeds, the conversational participants take turns at being speakers and hearers and the result both speaker and hearer have imperfect information about each other but also learn about each other's type. Thus ME games are much more general than signaling games. However, note that it is simple to model signaling games in the setting of ME games – the Jury declares the winner after the first two turns of the ME game.

5 Conclusion

In adding notions from epistemic game theory to ME games, we have shown how types account for certain conversational strategies and how they influence interpretation. We've shown that players strategize about their options given the types they assign to the Jury and to each other. We've also seen how the Jury updates its assessment of player types based on their contributions and determines whether the play is winning or losing for one of the players. Such reassessments of types also affect the interpretation of a discourse move and in the fashioning of a history out of an underspecified play. This allowed us to explain why bias is a natural outcome of interpretation and how it crucially depends on the underspecification of discourse connections between speakers' contributions. Our use of formal theories of discourse interpretation as providing the vocabulary of ME games provides a much more nuanced view of subjectivity in interpretation than has been proposed using signaling games [11].

Types are basically a device for assigning probabilities to other types, and ultimately to strategies, which in ME games are a function of the probabilities assigned to the types of the other players and of the Jury. We've shown that types furnish a linguistically unstudied but important component of interpretation. Even the construction of a logical form for a conversation is subject to type of the interpreter, which is common sensical enough but is something which linguistic theories have largely ignored.

Griceans and Neo-Griceans argue that speaker intentions are crucial to interpretation; but this misconstrues the contribution of epistemic information. Interpreters don't have access to the intentions and beliefs of the speaker, they have only their own beliefs about the speaker that interact with conventional cues for linguistic meaning. What is crucial for interpretation are the uncertain beliefs that interpreters have about the speaker, beliefs about the speaker's beliefs, which include the speaker's beliefs about the interpreters, her goals, and what she wants to do. Types encode this information in a formally precise framework.

Parametrizing SDRS construction and discourse interpretation to types has has several interesting consequences. The first is that it generalizes and formalizes an idea from [13, 17] that different conversational participants may construct different SDRSs for a given dialogue that nevertheless share some structure. This also makes a difference to commitments as [17] explain; in saying something speaker 0 may take herself to commit to p but player 1 may take 0 to commit to q, which may then be the basis for what 1 contributes next. Parametrizing interpretation relative to types while keeping basic meaning constant, means that any two such SDRSs will, assuming no processing errors, share the set of EDUs but may differ on how these are related or combined into larger CDUs. Of course two interlocutors might both have the same type and then would perforce construct the same SDRS. Parametrizing interpretation relative to types also predicts that what interpretation results may shift as we consider higher order beliefs about types, something not considered in other work in semantics or discourse, as far as we know. Finally, this mechanism predicts that as probabilities assigned to types

are updated, interlocutors may revise their interpretation of the contributions of their conversational partners, another intuitively compelling point.

A second consequence of this view opens up some intriguing generalizations about how linguistic interpretation can confirm expectations. As we saw in case iv with the biased Jury for the example about Sheehan, since types help determine the interpretation of an exchange and that exchange in turn helps one update ones' views about the types of the interlocutors, assigning a high probability to a particular type can be further confirmed by how it influences or even "writes" the discourse history. This predicts that biased Juries may hear "only what they want to hear" and that they do not change their minds even in the face of evidence that would convince an unbiased Jury.

The third consequence of this view is that people with different beliefs or even different moods (e.g. the interpreter is very angry or very suspicious) may produce different interpretations of a conversation. Different people, or the same person in different moods, can interpret the same verbal signal differently, though the exact parameters of variation would have to be (and perhaps already has been) made precise and empirically tested. Types can also be used to encode the variation in conversational interpretation between normal interpreters and neurally or cognitively impaired people. This opens up a large space of investigation that we leave for future research.

The use of types in interpretation also raises some important questions. What types are relevant to interpretation? The space of types is itself vast and raises technical difficulties for the existence of optimal strategies [6], but our intuitions tell us that typically the types relevant to a particular conversation are rather few. A fair interpretation imposes constraints on the set of relevant types, but at present we do not understand how this set is picked out.

We leave open several extensions to this work. The first concerns the notion of a fair conversational setting. Conversations, esp. strategic ones, if allowed to continue in an unrestricted fashion, may become unfair. For instance, it might be the case that the player i who gets the opportunity to speak first might not concede the turn to the other player(s) $(1-i)$ thus having an unfair advantage. Or i might talk over $(1-i)$ not allowing her to have her turn.[4] Conversationalists are aware, at least implicitly, of the dangers of such cases and debates have exogenous means of ensuring that there are optimal strategies for the speakers to follow. For instance, in debates there is usually a 'moderator' who ensures that all the participants get a fair chance to speak. She might interrupt a speaker and pass the turn on to another speaker. We could add a moderator entity to our existing model of ME games. A passive moderator would have the responsibility of the moderator is to assign turns to the players and could esure that no player monopolized the conversation and that each player had a chance to respond to an interlocutor's assertion. We can also imagine active moderators and active Juries that actually participate at least in some minimal way in the conversation. We hope to tackle these issues in future research.

[4] See [6] for a case where an unrestricted conversation might assign players victory conditions with no pair of equilibrium strategies.

6 Appendix

6.1 The Cantor topology for finite and infinite strings

For any subset A of X, as usual, we denote by A^* the set of finite strings over A and by A^ω, the set of countably infinite strings over A. Let $\epsilon \in A^*$ be the empty string and let $A^\infty = (A^* \cup A^\omega)$. We define a metric d on X^ω as follows. Let $x = x_0x_1\ldots$ and $y = y_0y_1\ldots$ be infinite sequences in X^ω where each $x_j, y_j \in X$, $j \geq 1$. Let $n(x,y)$ be the first index where x and y differ. That is, $x_{n(x,y)} \neq y_{n(x,y)}$ and $x_j = y_j$ for all $0 \leq j < n(x,y)$. Then

$$d(x,y) = \frac{1}{2^{n(x,y)}}$$

d generates a (complete) metric space on X^ω usually known as the Cantor topology on X^ω. We extend the metric d to X^∞ by letting $\$ \notin X$ be a new symbol, identifying every $x \in X^+$ with $x\$^\omega$ and extending the metric d to $(X^\omega \cup X^+\$^\omega)$. Under this metric, X^+ gets the discrete topology (every subset is both open and closed) and every open set in X^∞ is of the form $(A \cup BX^\omega)$ where $A, B \subset X^+$. Moreover, (X^∞, d) has the nice property of being the completion of the metric space (X^+, d) (see [15] for more details). Given any non-empty set X, we shall work with Borel sigma algebras over X^+, X^ω and X^∞ that are generated by the topology defined above. Unless otherwise mentioned, a product of topological spaces will be assigned the product topology. Given any set X we shall denote its relative complement by $\overline{X}$, when the universe is clear from the context.

6.2 Higher order beliefs

Definition 12 (Event). *An* event E_i^ρ *at* $\rho \in \mathcal{P}$ *for Player* i *or an* i-event *is a measurable subset of* $(T_{(1-i)} \times S^\rho_{(1-i)} \times T_{\mathcal{J}})$. *Similarly, an event* $E^\rho_{\mathcal{J}}$ *for the Jury or a* $\mathcal{J}$-event *is a measurable subset of* $(T_0 \times S_0^\rho \times T_1 \times S_1^\rho)$.

$\beta_i^\rho(t_i, \mathsf{h})(E_i^\rho)$ is the subjective probability that Player i assigns to E_i^ρ given history $\mathsf{h} \in \mathsf{h}(\rho)$ and given that she is of type t_i. Accordingly, t_i is said to believe E_i^ρ at history h if $\beta_i^\rho(t_i, \mathsf{h})(E_i^\rho) = 1$. Furthermore, the subjective probability that Player i assigns to E_i^ρ given play ρ is denoted by $\beta_i^\rho[t_i](E_i^\rho)$ and is defined as:

$$\beta_i^\rho[t_i](E_i^\rho) = \int_{\mathsf{h} \in \mathsf{h}(\rho)} \xi_i(t_i, \rho)(\mathsf{h}) \beta_i^\rho(t_i, \mathsf{h})(E_i^\rho) d\xi_i(t_i, \rho)$$

t_i is said to believe E_i^ρ at play ρ if $\beta_i^\rho[t_i](E_i^\rho) = 1$. The Bayesian conditional beliefs are computed as usual. For example, given i-events E_i^ρ and F_i^ρ at ρ, where F_i^ρ is assumed to be non-null:

$$\beta_i^\rho(t_i, \mathsf{h})(E_i^\rho | F_i^\rho) = \frac{\beta_i^\rho(t_i, \mathsf{h})(E_i^\rho \cap F_i^\rho)}{\beta_i^\rho(t_i, \mathsf{h})(F_i^\rho)}$$

$$\beta_i^\rho[t_i](E_i^\rho | F_i^\rho) = \frac{\beta_i^\rho[t_i](E_i^\rho \cap F_i^\rho)}{\beta_i^\rho[t_i^\rho](F_i^\rho)}$$

Type t_i of Player i believes E_i^ρ at ρ given F_i^ρ if $\beta_i^\rho[t_i](E_i^\rho|F_i^\rho) = 1$.

For an i-event E_i^ρ at ρ, let

$$B_i^\rho(E_i^\rho)^{(1-i)} = \{(t_i, \sigma_i, t_\mathcal{J}) \in (T_i \times S_i^\rho \times T_\mathcal{J}) \mid t_i \text{ believes } E_i^\rho\}$$

$$B_i^\rho(E_i^\rho)^\mathcal{J} = \{(t_0, \sigma_0, t_1, \sigma_1) \in (T_0 \times S_0^\rho \times T_1 \times S_1^\rho) \mid t_i \text{ believes } E_i^\rho\}$$

Note that $B_i^\rho(E_i^\rho)^{(1-i)}$ is an $(1-i)$-event and $B_i^\rho(E_i^\rho)^\mathcal{J}$ is a $\mathcal{J}$-event. Then type $t_{(1-i)}$ of Player $(1-i)$ believes that Player i believes E_i at ρ if $\beta_{(1-i)}^\rho[t_{(1-i)}](B_i^\rho(E_i^\rho)^{(1-i)}) = 1$ and type $t_\mathcal{J}$ of the Jury believes that Player i believes E_i at ρ if $\beta_\mathcal{J}^\rho[t_\mathcal{J}](B_i^\rho(E_i^\rho)^\mathcal{J}) = 1$.

Similarly, for a $\mathcal{J}$-event $E_\mathcal{J}^\rho$, let

$$B_\mathcal{J}^\rho(E_\mathcal{J}^\rho)^i = \{(t_{(1-i)}, \sigma_{(1-i)}, t_\mathcal{J}) \in (T_{(1-i)} \times S_{(1-i)}^\rho \times T_\mathcal{J}) \mid t_\mathcal{J} \text{ believes } E_\mathcal{J}^\rho\}$$

That is, $B_\mathcal{J}^\rho(E_\mathcal{J}^\rho)^i$ is an i-event. Then type t_i of Player i believes that the Jury believes $E_\mathcal{J}^\rho$ at ρ if $\beta_i^\rho[t_i](B_\mathcal{J}^\rho(E_\mathcal{J}^\rho)^i) = 1$. These are the second-order beliefs of the players and the Jury. Continuing this way, we can define any (finite) level of higher-order belief of the players, the Jury, the higher order beliefs of the players about the Jury and vice-versa etc.

6.3 Expected Utilities

We define expected utilities inductively.

Let ρ be a play and t_i be a type of Player i. Define the following update operation.

– $$u_{t_i}^0(\mathsf{h}_j) = \begin{cases} 1 & \text{if } \mathsf{h} \in \mathsf{h}(\rho) \text{ and } \mathsf{h}_j \notin \mathit{Win}_{(1-i)} \\ 0 & \text{otherwise} \end{cases}$$

where $0 \leq j \leq |\rho|$.

– For every history $\mathsf{h} \in \mathsf{h}(\rho)$ and for every prefix h_j of h where $0 \leq j < |\rho|$ define the 1-turn extension of h_j as

$$\mathbf{h}_j = \{\mathsf{h}' \mid |\mathsf{h}'| = j+1 \text{ and } \mathsf{h}_j \text{ is a prefix of } \mathsf{h}'\}$$

Then $u_{t_i}^{k+1}(\mathsf{h}_j)$ for $k \geq 0$ is given by:

$$\sum_{\mathsf{h}' \in \mathbf{h}_j} \int_{(T_{(1-i)} \times S_{(1-i)}^{\rho_{j+1}} \times T_\mathcal{J})} \beta_i^{\rho_{j+1}}(t_i, \mathsf{h}')(t_{(1-i)}, \sigma_{(1-i)}, t_\mathcal{J}) \xi_i^{\rho_{j+1}}(t_i, t_{(1-i)}, t_\mathcal{J})(\mathsf{h}') u_{t_i}^k(\mathsf{h}') d(t_i, t_{(1-i)}, t_\mathcal{J})$$

The expected utility is then defined as the limit of the above operation as $k \to \omega$ for the history h_0 and is denoted as $u_{t_i}(\rho)$.

Definition 13 (Expected utility). *The expected utility for the type t_i of Player i for the play ρ is given by $u_{t_i}(\rho) = u_{t_i}^\omega(\mathsf{h}_0)$.*

Recall that although some plays are finite, they are in *Win* only if all their continuations are. It is then straightforward to show that $u^{\omega}_{t_i}(\mathsf{h}_j)$ exists and is unique for all $j \geq 0$ meaning that the expected utility is well-defined.

Lemma 1. *Given a play ρ of the ME game, for every type t_i of Player i, and for every $j \geq 0$, $u^{\omega}_{t_i}(\mathsf{h}_j)$ exists and is unique.*

Proof. Simply note that for every $j \geq 0$, $\{u^{k}_{t_i}(\mathsf{h}_j)\}_{k\geq 0}$ forms a non-increasing Cauchy sequence.

6.4 Expected payoffs for strategies

Definition 14 (Expected payoff). *Let $\mathcal{G}$ be an ME game and let $p \in \Delta(S_{(1-i)})$ be a probability measure over the set of strategies of $(1-i)$. The* expected payoff *of a strategy $\sigma_i \in S_i$ of type t_i of Player i with respect to p is defined as the Lebesgue integral:*

$$u_{t_i}(\sigma_i, p) = \int_{\sigma_{(1-i)} \in S_{(1-i)}} u_{t_i}(\rho_{(\sigma_i, \sigma_{(1-i)})}) dp$$

σ_i maximizes expected payoff *for type t_i of Player i with respect to p provided for all $\sigma'_i \in S_i$, $u_{t_i}(\sigma_i, p) \geq u_{t_i}(\sigma'_i, p)$. Then, we also say that σ_i is a* best response *to p for type t_i.*

Let $\mathcal{T} = (\{T_i\}, T_{\mathcal{J}}, \{\hat{\beta}^{\rho}_i\}, \{\hat{\beta}^{\rho}_{\mathcal{J}}\}, S)$ be a type space for an ME game $\mathcal{G}$. Recall that, given a play ρ, for every history $\mathsf{h} \in \mathsf{h}(\rho)$, each type $t_i \in T_i$ is associated with a probability measure $\beta^{\rho}_i(t_i, \mathsf{h}) \in \Delta(T_{(1-i)} \times S^{\rho}_{(1-i)} \times T_{\mathcal{J}})$. Then, for each t_i, define a probability measure $p^{\rho}_{t_i} \in \Delta(S_{(1-i)})$ as: for every measurable subset $S'_{(1-i)}$ of $S_{(1-i)}$ let

$$p^{\rho}_{t_i}(S'_{(1-i)}) = \sum_{\mathsf{h} \in \mathsf{h}(\rho)} \int_{(T_{(1-i)} \times T_{\mathcal{J}})} \xi^{\rho}_i(t_i, t_{(1-i)}, t_{\mathcal{J}})(\mathsf{h}) \cdot \beta^{\rho}_i(t_i, \mathsf{h})(t_{(1-i)} \times S'_{(1-i)} \times t_{\mathcal{J}}) d(t_{(1-i)}, t_{\mathcal{J}})$$

References

1. S. Afantenos, E. Kow, N. Asher, and J. Perret. Discourse parsing for multi-party chat dialogues. In *Empirical Methods in Natural Language Processing*, pages 928–937. Association for Computational Linguistics, 2015.
2. N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.
3. N. Asher and A. Lascarides. Strategic conversation. *Semantics and Pragmatics*, 6(2):http:// dx.doi.org/10.3765/sp.6.2., 2013.
4. N. Asher and S. Paul. Evaluating conversational success: Weighted message exchange games. In J. Hunter, M. Simons, and M. Stone, editors, *20th workshop on the semantics and pragmatics of dialogue (SEMDIAL)*, New Jersey, USA, July 2016.

5. N. Asher, S. Paul, and A. Venant. Message exchange games in strategic conversation. *Journal of Philosophical Logic*, 2015. In press.
6. Nicholas Asher and Soumya Paul. Language games. In *Proceedings of the 20th International Conference on Logical Aspects of Computational Linguistics (LACL)*, LNCS, Berlin Heidelberg, 2016. Springer. to appear.
7. Nicholas Asher and Soumya Paul. Strategic reasoning in conversations under imperfect information. In Z. Sikić et al, editor, *Proceedings of the 5th Conference on Logic and Applications (LAP)*, Dubrovnik, 2016. IUC.
8. Robert Aumann. Borel structures for function spaces. *Illinois Journal of Mathematics*, 5:614–630, 1961.
9. Eddie Dekel and Marciano Siniscalchi. *Epistemic Game Theory*, pages 619–702. 2015.
10. Juan Dubra and Federico Echenique. Information is not about measurability. *Mathematical Social Sciences*, 47(2):177–185, 2004.
11. M. Franke. Meaning and inference in case of conflict. In Kata Balogh, editor, *Proceedings of the 13th ESSLLI Student Session*, pages 65–74, 2008.
12. John C Harsanyi. Games with incomplete information played by bayesian players, parts i-iii. *Management science*, 14:159–182, 1967.
13. A. Lascarides and N. Asher. Agreement, disputes and commitment in dialogue. *Journal of Semantics*, 26(2):109–158, 2009.
14. Eric Pacuit and Olivier Roy. Epistemic foundations of game theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, 2015.
15. D. Perrin and J. E. Pin. *Infinite Words - Automata, Semigroups, Logic and Games*. Elsevier, 1995.
16. L.M. Solan and P.M. Tiersma. *Speaking of Crime: The Language of Criminal Justice*. University of Chicago Press, Chicago, IL, 2005.
17. A. Venant, N. Asher, and C. Degremont. Credibility and its attacks. In V. Rieser and P. Muller, editors, *The 18th Workshop on the Semantics and Pragmatics of Dialogue*, pages 154–162, 2014.