Bias in Semantic and Discourse Interpretation

Nicholas Asher, Julie Hunter, Soumya Paul

Bias in Semantic and Discourse Interpretation*

Nicholas Asher[1], Julie Hunter[2] and Soumya Paul[3]

[1] CNRS, IRIT, Université Paul Sabatier, Toulouse, France
nicholas.asher@irit.fr
[2] LINAGORA
jhunter@linagora.com
[3] CSC, FSTC, Université du Luxembourg, Esch-sur-Alzette, Luxembourg
soumya.paul@gmail.com

Abstract. In this paper, we show how game-theoretic work on conversation combined with a theory of discourse structure provides a framework for studying interpretive bias and how bias affects the production and interpretation of linguistic content. We model the influence of author bias on the discourse content and structure of the author's linguistic production and the influence of interpreter bias on the interpretation of ambiguous or underspecified elements of that content and structure. Interpretive bias is an essential feature of learning and understanding but also something that can be exploited to pervert or subvert the truth. We develop three types of games to understand and to analyze a range of interpretive biases, the factors that contribute to them, and their strategic effects.

1 Introduction

Bias is generally considered to be a negative term, but bias is in fact essential to understanding [47]: one cannot interpret or make sense of data—something humans are disposed to try to do even when that data is nothing but noise [35]—without relying on a bias or hypothesis to guide that interpretation.[4] Yet biases can lead to very different understandings of what objectively or naively we would take to be the same event. This paper provides a formal account of how biases on the part of the author and of an interpreter affect linguistic interpretation. In particular, we model the influence of author bias on the author's discourse's content and its structure and interpreter bias on the interpretation of ambiguous or underspecified elements of that content and structure. We investigate how bias affects linguistic content not only when authorial and interpreter bias are aligned but also when they are opposed or incompatible and then use this to model how agents can interpret intuitively the same event or text quite differently. We develop our analysis of bias within a game theoretic framework, and we present three different types of games to isolate strategic uses of bias, linguistic indicators of bias, and of how biases determine whether a conversational participant can meet her conversational goals.

* We thank the ANR PRCI grant SLANT and the 3IA Institute ANITI funded by the ANR-19-PI3A-0004 grant for research support. We thank anonymous reviewers from *Linguistics and Philosophy* for extensive and helpful comments on a previous draft .

[4] This point can be made mathematically precise [66, 65].

One might think of biases as simply prior beliefs, in which case, our puzzle about biases might seem to amount to an uninteresting truism: what we come to believe from our observations is colored by our prior beliefs. If biases are beliefs, they are deep seated ones and beliefs of a particular kind; they are often inaccessible or at least not easily accessible to our conscious thought processes. But even if biases are just beliefs, most current formal semantic models do not take bias into account. In formal semantics and pragmatics, we generally investigate devices—words, prosodic contours, phrasal constructions, discourse structures—by which intersubjectively stable meaning gets conveyed. From this traditional perspective, it's not clear where biases can intrude in the meaning construction process nor how they could or why they could shape it. The formal, model theoretic tradition in semantics and many pragmatic accounts of meaning since Frege doesn't leave a place for biases or beliefs to get in to influence meaning.

A place to start is with the following observation. The way in which an author structures a text or discourse, what discourse moves she makes, should rationally depend on her particular bias and on what she wants to achieve, and that in turn depends on what she believes the bias of her readers to be. Similarly though less obviously, a reader's bias will take into account what she estimates the author's intentions and bias to be. And this now invites the question, when does a choice of a sequence of discourse or conversational moves, given the biases of the author and the interpreter, lead to conversational success in which both achieve their aims, in which only one participant achieves their aims, and in which none of the participants do? We are especially interested in this question when authorial and interpreter biases are incompatible.

This naturally suggests a game theoretic framework for analyzing bias. But while signalling games [44, 61, 26] have shown how the transfer of intersubjective meaning results from a cooperative coordination between speaker and hearer, they have problems of interpretability when assumptions of cooperativity and other cognitive hypotheses are not met [62]—assumptions that we cannot make in this study. In addition, signalling game approaches are not designed to distinguish between intersubjectively stable meaning and those underspecified elements that different biases can endow with different content, as the whole aim of signalling games is to give an account of coordinated intersubjective meaning.[5] We need a theory of interpretive bias that starts from linguistically given and grammatically determined content and introduces authorial and interpreters' biases that may lead to cooperativity and coordination, but also to misunderstanding, manipulation and deceit.

[9, 2, 8] develop a general account of the strategic aims and consequences of conversations in a framework called *Message Exchange (ME) Games* that will provide the foundations of our novel analysis of interpretive bias for language. ME games build on an account of intersubjective meaning provided by theories of discourse structure and interpretation like Segmented Discourse Representation Theory or SDRT [3, 5], although they are designed to analyze the interaction between conversational structure, purposes, assumptions and strategies, in the absence of assumptions about cooperativity or other cognitive hypotheses. ME games define conversational goals precisely as winning conditions in an ME game, and thus will will furnish what we need to define biases. In addition, their exploitation of discourse theoretic notions of structure and

[5] See also [49–51] in this vein.

meaning will allow us to investigate how bias helps build meaning in conjunction with discourse or conversational moves and how that meaning can in turn convey and affect biases, as we will see informally in Section 2 and then more formally in Section 3. Section 3 will also develop the interaction of beliefs and interpretations, when we introduce the essentials of ME games and of SDRT.

In Section 4, we introduce three new types of ME games to analyze biases and their strategic uses. The first sort of game, an *ME evaluation game* will show a relatively simple strategic use of bias in pursuit of the straightforward conversational goal of having the interpreter won over by the author's contribution. We will be able to see how aligned biases lead to the accomplishment of this goal, and how opposed biases can make accomplishing this goal very difficult. In addition we will show how the interactions of beliefs and interpretations lead very naturally to a reinforcement of biases, which in turn affect conversational success and which also in turn leads to a by now familiar hostility to alternative points of view. ME evaluation games will also allow us to analyze *epistemic contents*, components of discourse content that depend upon the biases of the game's participants. Nevertheless, ME evaluation games are too simple to evaluate the strategic uses of epistemic content, and so we introduce a more complex ME game, known as an *ME persuasion game*. They will further allow us to model different types of interpretive bias with different goals besides the usual scientific goals of accurate prediction, and this leads naturally to a discussion of optimal interpretive biases when ground truth is not independently accessible—something we analyze with another type of ME game, *ME truth games*. This last point is especially important in light of the problem of bias reinforcement mentioned above.

2 Preliminaries on interpretive bias

As we've suggested, biases that affect linguistically conveyed content can depend on authorial biases as well as interpreter biases. Authorial biases create different narratives or representations of what we would intuitively consider to be the same sequence or collection of events through choices of how to describe events and importantly *which events* in that collection to describe. Consider the lead paragraphs of articles from the *New York Times*, *Townhall* and *Newsbusters*, concerning the March for Science held in April, 2017. They are segmented into roughly clausal level units or *discourse units*—so called because they are the building blocks of discourse structure.

(1) (a) The March for Science on April 22 may or may not accomplish the goals set out by its organizers. (b) But it has required many people who work in a variety of scientific fields — as well as Americans who are passionate about science — to grapple with the proper role of science in our civic life. (c) The discussion was evident in thousands of responses submitted to NYTimes.com ahead of the march, both from those who will attend and those who are sitting it out.

(2) (a) Do you have march fatigue yet? (b) Well the left, apparently, does not, (c) so we're in for some street theater on Earth Day, April 22, with the so-called March for Science. (d) It's hard to think of a better way to undermine the public's faith in science than to stage demonstrations in Washington, D.C., and around the country modeled on the Women's March on Washington that took place in January. (e) The Women's March was an anti-

Donald Trump festival. (f) Fine. (g) I found it vulgar and demeaning to women, (h) but it's a free country. (j) Science, however, to be respected, must be purely the search for truth. (k) The organizers of this "March for Science" – (l) by acknowledging that their demonstration is modeled on the Women's March – are contributing to the politicization of science, (m) exactly what true upholders of science should be at pains to avoid.

(3) (a) Thousands of people have expressed interest in attending the "March for Science" this Earth Day, (b) but internally the event was fraught with conflict (d) and many actual scientists rejected the march and refused to participate.

These different articles begin with some of the same basic facts: the date and purpose of the march, and the fact that the march's import for the science community is controversial, for example. But the three texts convey very different pictures. (1), for instance, interprets the controversy as generating a serious discussion about "the proper role of science in our civic life," while (2) interprets the march as a political stunt; it characterizes the march for science as *street theater*, as a demonstration like a political protest. (3) also paints a more negative picture of the march.

While the choice of wording clearly can convey bias,[6] just as crucial is which events authors choose to include in their narrative and which they leave out, something as far as we know hasn't been examined. (2)'s bias against the March of Science expressed in the argument that it politicizes science cannot be traced back to negative opinion words applying to the March itself. Each article colors the event by leaving out certain information. For instance, there is no discussion that serious scientists have been involved in (2)'s account. On the other hand, (1)'s description does not mention any demonstrations but mentions discussions between scientists and the general public.

A third and largely unexplored way in which authors convey bias is to use rhetorical structure to link the events they describe into a coherent narrative. The *Townhall* article starts out with a rhetorical question and then supplies an answer to the question from the "Left" that it then comments on negatively. Given this characterization, one can discern a discourse goal of the author to appeal to skeptics of the March for Science with more right wing ideologies, those who would not be in favor of "marches," "street theater," and "demonstrations." In particular the ways in which the march is related to other events described are crucial factors in conveying bias. Once again the *Townhall* article relies on a comparison (in units (d) and (e)) between the March for Science and the Women's March, which is portrayed as a political, anti-Trump event. By implication the March for Science is also characterized as such. *Newsbusters* paints a somewhat negative view of the March via a discourse connection of Contrast—between the interest many people had in the march and the assertion that many actual scientists rejected it. The use of Contrast on the contents in this order is crucial; as [16] show, the second terms of a Contrast relation used in discourse theories like SDRT or RST [45, 3, 5] convey the discourse point or main opinion of the story. Another instance of Contrast, occurs in the *Townhall* story; there is an implicit contrast between constituent *j* and the clauses *e* and *k*— between science as *purely the search for truth* and the political

[6] A considerable amount of research especially in computational linguistics has been done in this area; see [46, 43, 19]. We note also that [19] shows that restricting oneself to lexical biases paints a very imperfect picture of what bias is.

purpose of the Women's March—which once again reinforces the view of the March for Science, modelled on the Women's March (this is set up in constituent (d)), as not really about science. (1) also uses a contrast between its first and second sentences to convey a more favorable view of the event. In light of these uses of semantic or rhetorical relations, we would naturally infer a different discourse aim for (1)'s author from that for the *Townhall* author.

To sum up, we see that the basic components driving interpretive bias in language are: 1) the events or more generally propositions one chooses to attend to, 2) how they are described, and 3) the discourse or semantic relations that determine how the propositions explicitly introduced are linked together.

We've isolated some components of discourse that convey bias that an author can choose. What about interpreter bias? In many cases, authors choose to leave connections or descriptive content underspecified for a variety of purposes that we will discuss later in this paper, thus inviting the interpreter to use her biases to fill in the picture. Consider:

(4) The meeting has been cancelled. Julie didn't show up.

While these clauses are not explicitly connected with a particular semantic relation, an interpreter will typically have antecedent biases that lead her to interpret eventualities described by the two clauses as entering into one of two configurations: one in which the eventuality described by the first clause caused the second, or one in which the second caused the first. Though not explicitly introduced by the author, these semantic relations between the eventualities described in the text clearly contribute to its content. This is the contribution of interpreter's bias to linguistically determined content. Any time that such structural connections are left implicit by speakers—and this is much if not most of the time in text—interpreters will be left to infer these connections and thereby potentially create their own version of the events.

Taking stock from these observations, we see that a proper account of interpretive bias must attend to four elements: 1) what events the author chooses to talk about in her narrative, 2) how she describes the events she has chosen, using lexical choice, 3) how she relates them to each other in a narrative and 4) how her interpreter fills in those elements, including discourse structural elements, that the author leaves unsaid or unspecified. This means that our analysis of interpretive bias will need to appeal to a semantic theory in which discourse structure is a factor in determining content. Moreover, we will need to appeal to a semantic theory in which underspecified elements can become specified through an interaction of beliefs and grammatically determined contents. The theory of epistemic games developed in [8], which in turn builds on SDRT's theory of discourse semantics, details this interaction between beliefs and discourse content. We review briefly both SDRT and the relevant aspects of [8]'s epistemic games in the next section.

3 The tools we will use

In this section we describe the tools we need to build the formal model of interpretive bias. We first give an overview of how we will represent discourse content and structure using SDRT, as we have argued in the preceding section that taking account of discourse

structure is an important step in accounting for interpretive bias. We then describe the game theoretic framework we need.

3.1 SDRT: representing discourse structure and content

Segmented Discourse Representation Theory or SDRT [3, 5] is a theory of discourse structure and content. SDRT is a convenient tool for us because it makes use of underspecification but also fully specified representations of discourse contents. Such fully specified structures are known as *Segmented Discourse Representation Structures*, or SDRSs. An SDRS consists first of a set of *discourse units* (DUs) of a text, conversation or other information bearing structure, where a discourse unit is an instance of propositional content. The second component of an SDRS is a set of determinate semantic relations defined over those elements to form a coherent, weakly connected, acyclic graph with a unique root [5].[7] For example, take the first three sentences of (2), repeated here as (5):

(5) (a) Do you have march fatigue yet? (b) Well the left, apparently, does not, (c) so we're in for some street theater on Earth Day, April 22, with the so-called March for Science.

The graph in Figure 1 provides an example of a fully specified structure for (5).

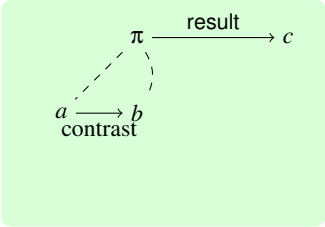

Fig. 1. Fully specified SDRS for the first three sentences of example (2)

Unit (a) expresses a rhetorical question; its presumed, positive answer stands in contrast to the unit (b)—with the contrast suggested by the discourse particle *well* (though the particle has other uses as well). As the answer to (a) is left implicit, we represent the Contrast relation with an arc from (a) to (b). Intuitively, (a) and (b) together describe the situation that the author is representing as the impetus for the plans for "street theater" described in (c)—you would only consider the march *street theater* if you have march fatigue and the left doesn't and so is planning to march. In Figure 1, we represent the fact that (a) and (b) jointly describe the cause for (c) by grouping

[7] That is, an SDRS must be a graph G with just one element that has no incoming arrows; in addition, there are no elements a, b of G such that the transitive closure of the arcs in G give us the arrows $a \rightarrow b$ and $a \leftarrow b$.

them into a *complex discourse unit* (CDU) π that then serves as the first argument to the Result relation that links to (c). SDRT allows for the construction of such complex units built from smaller units; hence, the graph representation of an SDRS typically has two types of arcs, one (represented with a solid arrow in Figure 1) representing semantic relations between DUs that may get typed as instantiating a specifc semantic relation and another (represented with dotted lines in Figure 1) representing the constitution of complex discourse units.

While ideally the objective clues provided by the author should determine a fully specified SDRS, in reality, texts and conversations often fail to do so via the grammar of the language alone. In (c) for example, the discourse marker *so* indicates that a Result relation is at work, but nothing in the grammar specifies the cause, or the first argument of this relation. The grammar typically produces an *underspecified logical form* or *ulf*, that is, a structure with elements, typically relation instances, arcs, or their labels (what semantic relations they represent) but also basic elements of the graphs, that are underspecified.

To make these ideas more precise, we review SDRT's formal language, which we call V here. The vocabulary of V contains a countable distinguished set of individual constants or discourse unit labels $\mathsf{DU} = \{\pi, \pi_1, \pi_2, \ldots\}$, and a finite set of discourse relation symbols $\mathbb{R} = \{\mathcal{R}, \mathcal{R}_1, \ldots \mathcal{R}_u\}$, as well as a countable set of variables for DUs and relations. Discourse unit labels tag 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 DUs—like the relation of Contrast (`contrast`) or the relation of *Question Answer Pair* (`qap`), in which one move answers a prior move characterized by a question. Finally, V contains formulas $\phi, \phi_1, \ldots$ from some fixed language $\mathcal{L}$ for describing elementary discourse move contents, a language like that of higher order logic used in, e.g., Montague Grammar. The SDRT formulas of V are of the form $\langle \pi : \phi \rangle$, where ϕ is either: (i) a formula of $\mathcal{L}$, (ii) 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, (iii) conjunctions of SDRT formulas, or (iv) existential closures of such formulas. When ϕ is a formula of $\mathcal{L}$, 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.

We now illustrate our formal language with our *Townhall* example. The formal representation for (5)a, (5)b and (5)c would be $\pi_a : \phi_a$, $\pi_b : \phi_b$, $\pi_c : \phi_c$. The DU ϕ_c bears a clear semantic relation, the relation *Result*: ϕ_c is the result of some event or events described in prior discourse units. 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 cause of ϕ_c could be the event described in ϕ_b or the complex of events described in the complex discourse unit consisting of ϕ_a and ϕ_b. Since we don't know, we say the first element of the Result relation, the cause, is underspecified, and we existentially quantify over a DU variable to signify this, as shown in (6). In addition, the relation between π_a and π_b is also not clear; the grammar tells us

that π_a is connected to π_b 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_b's furnishing an answer of the event described in π_a—something we denote in V by $\mathsf{qap}(\pi_a, \pi_b)$. Or we could interpret the relation as π_b's providing a contrast with an implied answer to the question ϕ_a (denoted in V by $\mathsf{contrast}(\pi_a, \pi_b)$). But because these relations have incompatible semantic preconditions (Contrast must have two arguments that are both true, while QAP requires its first argument to have a question which does not take a truth value at a world of evaluation), both cannot apply to π_a and π_b. In such a case, we say that the relation between π_a and π_b is *underspecified* and we simply quantify over a relational variable to signify this as shown in (6). So we can now provide the formula in V representing the grammatically determined content of 5a-5c:

(6) $\quad 5a5b5c \equiv (\langle \pi_a : \phi_a \rangle \wedge \langle \pi_b : \phi_b \rangle \wedge \exists \mathcal{R} \cdot \mathcal{R}(\pi_a, \pi_b) \wedge \langle \pi_c : \phi_c \rangle \wedge \exists \pi \cdot \mathsf{res}(\pi, \pi_c))$

The formula on the right hand side of $\equiv$ in (6) is what [5] calls an *underspecified logical form* or ULF. The relation between π_a and π_b is underspecified, because the grammar (syntax, compositional and lexical semantics) does not determine it. (6) corresponds to the graph in Figure 2. Once the existentially quantified variables in (6) are instantiated,

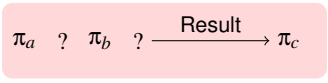

Fig. 2. Underspecified SDRS for example (5)

we get a *fully specified logical form* or FLF. A single ULF can give rise to a (finite) number of FLFs or fully specified SDRSs, as, in the language of SDRT, an FLF is nothing but a complete SDRS. The FLF corresponding to the graph in Figure 1 is:

(7) $\quad (\langle \pi : (\langle \pi_a : \phi_a \rangle \wedge \langle \pi_b : \phi_b \rangle \wedge \mathsf{contrast}(\pi_a, \pi_b)) \wedge \langle \pi_c : \phi_c \rangle \wedge \mathsf{res}(\pi, \pi_c)) \rangle)$

3.2 Basics of ME games

We now move to the main component of our analysis of interpretative ME games. First, we revisit some of the definitions for ME games introduced in [8] and [9]. We then introduce the notions of types and belief functions and the tools required to develop them in our setting. Our ME games will assume two players (conversationalists), 0 and 1. While the technical definitions below can be easily lifted to the case where there are more than two players, multiparty conversation is substantially different from and more complex than two party conversations; a formal generalization of our theory of bias to multiparty conversations is beyond the scope of this paper. In what follows i will always take a value in $\{0,1\}$. Player $(1-i)$ will denote the opponent of player i.

[8] defines an ME game as a finite or infinite game over V, our formalization of the SDRT language introduced above. 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 is 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 the set $(V_0 \cup V_1)$. The set of all possible conversations 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. This union is $((V_0 \cup V_1)^* \cup (V_0 \cup V_1)^\omega)$, or $(V_0 \cup V_1)^\infty$. The set of all finite but non-empty sequences over $(V_0 \cup V_1)$ is denoted as $(V_0 \cup V_1)^+$.

While conversations are never actually infinite, it is sometimes strategically important to act as if they were, as [9] argues; and thus the infinitary character of ME games is useful for modeling certain strategic aspects of conversation. Moreover, we will use ME games as a theoretical tool to investigate biases, and some instances of the games we use, in particular ME persuasion games or ME truth games in Section 4, may make use of the potential infinitary character of ME games to describe certain types of biases.

We now formally define an ME game. In the definition, we use a term $\mathcal{J}$ which stands for a 'Jury' of the ME game. The Jury determines which player (or players) has achieved her goal in the conversation; in other words, it fixes the winning conditions in an objective fashion for the players. The Jury is typically an agent distinct from the players 0 and 1 of a ME game, but we can also sometimes identify the Jury with one of the players, something we exploit in Section 4.1. We define the Jury in Definition 4.

Definition 1 (ME game [9]). *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 move x_1, player 1 follows with x_2, player 0 then plays x_3 and so on.[8] This results in the sequence $x_1x_2x_3\ldots$. Given our language V that we developed in the previous section, this sequence is a concatenation of formulas from $V_0 \cup V_1$, where concatenation is viewed as conjunction. Consider the following conversation that extends example (4).

(8) a. ρ_1 = (*The meeting has been cancelled. Julie did not turn up.* 0)
b. ρ_2 = (*Do you know why she didn't show up?*, 1)

Assume player 0 plays the sequence ρ_1. ρ_1 yields a formula of V_0—a pair consisting of the V formula together with the index 0 for player 0: $[(\langle \pi_1 : \phi_1 \rangle \wedge \langle \pi_2 : \phi_2 \rangle \wedge \exists \mathcal{R} \cdot \mathcal{R}(\pi_1, \pi_2)), 0)]$, where π_1 and π_2 mark EDUs given by the two sentences in (8-a). Player 1 then plays the sequence ρ_2 which translates into a formula of V_1, itself a pair consisting of a formula in V for the EDU introduced by the question paired with 1. This results in the sequence $\rho_1\rho_2$. This motivates the following definition of a play of an ME game.

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

[8] SDRT provides SDRSs with a well defined relation of consequence as well as a notion of coherence [9, 7]. So we can define an equivalence relation $\sim$ on V based on the 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. A $\sim$ equivalence class of V is a *class* of discourse moves. Thus, when we talk of a 'move', we shall actually be referring to its class.

ρ can be a ULF like that for ρ_1 above. When ρ is a fully specified SDRS, we call it a history.

Definition 3 (History). *A* history $\mathfrak{h}$ *of an ME game is a play that is an* FLF *or equivalently a fully specified SDRS.*

Given a play ρ, $\mathcal{H}(\rho)$ 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. Clearly $\mathcal{H}(\rho)$ can contain multiple, distinct, even incompatible histories. For example there are at least two possible histories for the play ρ_1 in (8-a): one in which we instantiate the existential quantifier over relations to Explanation, meaning that π_b explains why π_a happened, and one in which we instantiate with Result, meaning that π_b was the result of the event in π_a.

(9) Histories for ρ_1 in (8)

a. $\mathfrak{h}_1(\rho_1) = [(\langle \pi_a : \phi_a \rangle \wedge \langle \pi_b : \phi_b \rangle \wedge \exp(\pi_a, \pi_b)), 0]$

b. $\mathfrak{h}_2(\rho_1) = [(\langle \pi_a : \phi_a \rangle \wedge \langle \pi_b : \phi_b \rangle \wedge \mathsf{res}(\pi_a, \pi_b)), 0]$

Let $|\rho|$ denote the number of turns in a play ρ and $|\mathcal{H}|$ denote the same for $\mathcal{H}$. We let $\mathcal{P}$ (resp. $\mathcal{H}$) denote the set of all plays (resp. histories), where $\varepsilon \in \mathcal{P}$ (resp. $\varepsilon \in \mathcal{H}$) is the empty play (resp. empty history).

We now formally define the concept of a Jury.

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.*[9] Win_i *is called the* Jury winning condition *or simply the* winning condition *for player i.*

Definition 5 (Winning plays/histories). *A play* ρ *(or history* $\mathfrak{h}$*) is said to be* winning *for player i if* $\rho \in Win_i$ *(or* $\mathfrak{h} \in Win_i$*).*

It might be that $(Win_0 \cap Win_1) \neq \emptyset$; then $x \in (Win_0 \cap Win_1)$ is winning for both players.

Players' strategies are an important element that players reason about. A strategy of player i tells us how i reacts to player $1-i$'s moves.

Definition 6 (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^+*, the finite positive sequences 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* $S = S_0 \times S_1$*.*

Let $\rho = x_0 x_1 \ldots$ be a play in an ME game where $x_0 = \varepsilon$ 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, $x_{j+1} = \sigma_i(\rho_j)$. Given a finite play ρ, we let S_i^ρ denote the set of all strategies σ_i of player i such that ρ conforms to σ_i and let S^ρ denote the set of all strategy pairs (σ_0, σ_1) such that ρ conforms to (σ_0, σ_1).

[9] In some cases it is important to impose a consistency constraint in the following sense for a winning condition: for every play ρ of the ME game, $\rho \in Win_i$ iff $\mathfrak{h}(\rho) \subset Win_i$. We do not do this here because some ME games we will explore in the next section feature a ULF as the contribution of one player, while the other provides an interpretation of the ULF. For other constraints see [8].

To see some examples of strategies, let's return to (8). Suppose 0 has played ρ_1; one strategy of 1 is to play a clarification question $\rho_{2'}$ like *did you mean that the meaning was cancelled because..?* to understand better which history $\mathfrak{h}_1(\rho_1)$ of (9-a) or $\mathfrak{h}_2(\rho_1)$ of (9-b) was intended. Another strategy is to assume that the intended history was (9-a) and to ask for an explanation of why she didn't show up. It is this latter strategy that conforms to the actual play in ρ_1, ρ_2 of (8).

Types, beliefs and interpretations We now turn to the epistemic component of ME games. Players' beliefs, or the subjective probabilities they assign to plays, moves, and strategies affect how they reason in an ME game, i.e. what they say or how they react to some conversational turn. And for this, a player's beliefs must include beliefs about other players's strategies and beliefs about them. This nested structure of higher order beliefs (beliefs about beliefs) can be expressed in different ways, but given that we are taking beliefs to be probability functions as is usual in game theory, a natural way to do this is to exploit the type of a player [30]. The type of a player i is a property of the player that encodes his behaviour, the way he strategizes, his personal biases, etc. The $i-types$ for a player i are the possible properties, possible behaviors relevant to the ME game, that i could instantiate. Rubrics like "right wing", "left wing" describe types that we will use in Section 4. We will assume probability distributions, written $\Delta(A)$, for sets of types or strategies A.

Definition 7 (Harsanyi type space [30]). *A Harsanyi* type space *for a set of strategies 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 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, how they affect what content they get from a message and how those messages affect their beliefs. 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 8 (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 functions*[10] $\hat{\beta}_i^\rho = (\beta_i^\rho, \xi_i^\rho)$ *where β_i^ρ is the* belief function *and ξ_i^ρ is the* interpretation function *defined as:*

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

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

The (first order) belief $\hat{\beta}_\mathcal{J}^\rho$ of the Jury is a similar pair of functions.[11]

[10] Because the set of strategies is uncountable, we need to restrict ourselves to measurable functions and measurable sets—for details see [8]. For our simple examples of finite plays, this restriction is satisfied, because they are all basic open sets in $(V_0 \cup V_1)$. For the infinitary games of section 4, matters are more delicate but we gloss over the details here.

[11] $\hat{\beta}_\mathcal{J}^\rho = (\beta_\mathcal{J}^\rho, \xi_\mathcal{J}^\rho)$ where the belief function β_i^ρ and the interpretation function ξ_i^ρ are defined as:

$$\beta_\mathcal{J}^\rho : T_\mathcal{J} \times \mathcal{H}(\rho) \to \Delta(T_0) \times \Delta(S_0^\rho) \times \Delta(T_1) \times \Delta(S_1^\rho)$$

$$\xi_\mathcal{J}^\rho : T_\mathcal{J} \times T_0 \times T_1 \to \Delta(\mathcal{H}(\rho))$$

Intuitively, by fixing a type for the players and the Jury, the respective interpretation function says how they interpret the current play; that is, what are the probabilities that they assign to each possible history arising from the current play.[12] The belief function returns the beliefs about the types and the strategies of the other players and/or the Jury given a history and a particular player type. Using the definitions of first order beliefs, S, the set of strategies, and types, we can define higher order beliefs, beliefs that players or the Jury have about the beliefs of other players (and the Jury) and fill out the epistemic picture of our players.[13]

In some cases, the beliefs or the interpretations of the players or the Jury may be independent of one or more components or those components may be fixed.[14] In that case we shall simply suppress those components. For example, player i's beliefs concerning the type of player $(1-i)$ and her strategies might be independent of what player i believes about the type of the Jury. In that case the belief of i is given by the function $\beta_i^{\rho} : T_i \times \mathcal{H}(\rho) \to \Delta(T_{1-i}) \times \Delta(S^{\rho}_{(1-i)})$. Similarly, when we talk about the interpretation function of i restricted to types t_0 and t_j, we can simply write: $\xi_i^{\rho} : T_i \to \Delta(\mathcal{H}(\rho))$. This will simplify our analyses of examples.

Let's once again go back to example (8) to see how types and interpretations might play out in a very simple scenario. Suppose we have two types for 0, roughly one, t_0^e according to which 0 intended to link π_b to π_a via the discourse relation of Explanation and another type t_0^r according to which 0 intended to link π_b to π_a via Result. Suppose 1 only has one type. In that case, the play ρ_1 together with $\beta_1^{\rho} : \mathcal{H}(\rho) \to \Delta(T_0)$ determines a probability distribution over the types for 0. In turn these types via $\xi_1^{\rho} : T_0 \to \Delta(\mathcal{H}(\rho))$ determine a probability distribution over the two histories (9-a) and (9-b) for player 1. [8] details how such distributions evolve as a conversation proceeds.

Putting β and ξ in Definition 8 together over their respective outputs reveals a correspondence between interpretations of plays and types for a fixed Jury type τ, which we call the *Types/History correspondence*.

Remark 1 The Types/History Correspondence: *In an ME game $\mathcal{G}$, every history yields a distribution over types for the players and every tuple of types for the players and the Jury fixes a distribution over histories.*

While β encodes a subjective element, because of the Types/History correspondence, the probability distributions for types and histories dynamically evolve under Bayesian updating with observations. Turning again to example (8), suppose initially 1's type t_1 assigns equal probabilities to t_0^e and t_0^r. This is a subjective belief. In this situation the proper continuation for 1 might be to ask a clarification question: *did you mean that the meeting was cancelled because Julie didn't show up?* In the case of an affirmative response, 1 will now pick $\mathfrak{h}_1(\rho_1)$ with very high probability. This in turn will strengthen

[12] 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.

[13] Winning conditions define a notion of utility and together with the belief functions of each player this yields a notion of expected utility. For the technical details, see [8].

[14] 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 $a_j, a'_j \in A_j$, $f(a_1, a_2, \ldots, a_j, \ldots, a_n) = f(a_1, a_2, \ldots, a'_j, \ldots, a_n)$.

his belief in or his bias towards t_0^e. We will see more examples of this bias strengthening (see also [8]) in the next section.

We now have the pieces to define our tool for analyzing our view of interpretive bias, which we describe with three hypotheses:

Definition 9. *An* Epistemic Message Exchange game (Epistemic ME game)*, $\mathcal{G}$, is an ME game, with a Harsanyi Type Space and belief functions for* 0*,* 1 *and the Jury, as defined in Definitions 7 and 8.*

- Hypothesis 1: we represent an interpretive bias as a probability distribution Δ over player and Jury types in an epistemic ME game $\mathcal{G}$; Δ determines player choices for describing events and for interpreting them and hence a distribution over histories.
- Hypothesis 2: Types consist of beliefs about winning conditions and strategies. Winning conditions encode preferences for certain histories, which may reflect exogenous beliefs (about political views, ethical principles, etc.).
- Hypothesis 3: Biases are dynamic and evolve through updating via the Types/History Correspondence.

4 Epistemic ME games for analyzing bias

In this section, we go beyond [8, 9], introducing novel types of epistemic ME games that will help us analyze interpretive bias and its effects. Each displays a different use of interpretive bias, and accordingly each is distinguished by a different winning condition. The first type of epistemic ME game is an ME evaluation game. In such a game the author, player 0, attempts to persuade the interpreter, player 1, of his point of view, while player 1 simply affirms or rejects the author's contributions. ME evaluation games show us how differing biases formally lead to the reinforcement of differing judgments and how authors' estimations of interpreter types affect their discursive strategies. ME evaluation games will also allow us to formally define what we call "epistemic content," which has received recent attention from linguists [31, 56, 36], and which player 0 can exploit to convey in effect two different histories. We then introduce ME persuasion games in which player 1 is an opponent who can question or attack player 0's proposed history. In ME persuasion games the Jury now plays the role of the evaluator and picks the winner in 0's and 1's debate. ME persuasion games will allow us to analyze strategic uses of epistemic contents and to see how biases can lead to histories proposed for many different purposes—to discover the truth or to understand, but also to conceal the truth, to praise or disparage, to persuade or to dissuade. These different purposes translate into winning strategies depending on the types of the interpreter and the Jury. Finally we turn to ME truth games, in which we fix the Jury type in an ME persuasion game so as to guarantee good epistemic practices. This allows us to investigate what biases can lead to the best approximation of the truth. But attaining this best approximation may, we show, be an infinitary goal, one only realized if we use the full power of an ME game.

4.1 ME evaluation games

Our first type of epistemic game assigns player 0 the role of author. She chooses a subset of facts from a set of data or facts X and makes a series of discourse moves, producing a

play ρ that is typically underspecified. Player 1, the audience, fills in any underspecified elements of ρ and at the end has constructed a history $\mathfrak{h}$ which is a finite sequence in V_1^*; he can then comment on $\mathfrak{h}$ with a "like" or a "dislike." These terms represent an ongoing evaluation of the history that the author has so far suggested to him with her moves in ρ. Below let $\alpha.\beta$ to represent the concatenation of α and β.

Definition 10. *An ME evaluation game is an epistemic ME game where the winning condition for* 0*, Win*$_0$*, is of the form* $((V_0 \cup V_1)^*.V_0^*.(V_1^*.\{\langle like \rangle\}))^\infty$*. In an ME evaluation game, we identify the Jury with player* 1*, who chooses Win*$_1$ *either to be Win*$_0$ *or to be of the form* $((V_0 \cup V_1)^*.V_1^*.\{\langle Nay \rangle\}))^\infty$.

In an ME evaluation game we identify the Jury with player 1, because player 1's beliefs determine the winning condition. The winning condition for 0 is that after a finite number of exchanges, player 1 always says "like" to the history he has built from successive prefixes of ρ; if ρ is finite, player 1 must end with a "like". Player 1's winning condition will involve specifying plays to construct a history; this may lead to a play that coincides with 0's winning condition or not. Thus, the plays in ME evaluation games will typically have a sequence of pairs each consisting of a ULF together with an FLF.

The strategies for 0 in an ME evaluation game involve a choice of which events to describe in her play as well as how to relate them with an eye to her discourse aims or, in our theory, her winning condition in the game. From the perspective of subjective rationality or rationalizability (an important criterion in epistemic game theory [14]), good biases for 0 in a conversation are those that lead to histories in the winning condition; bad biases lead to histories that do not achieve the winning condition. We have already illustrated such choices in our examples about the March for Science.

ME evaluation games reflect the fact that the authors of texts like (1), (2) and (3) typically have at least the discourse goal to please their readership, which we model here as another player. Various interpreter types will respond differently to the story as it unfolds, and so we will extend [8]'s analysis about Juries to show how interpreters' beliefs and discourse interpretations dynamically evolve. Hence, 0's beliefs about the types of 1 are crucial to her success and rationalizable behavior. ME evaluation games offer a simple setting where these constraints are in play.

To illustrate, we construct an ME evaluation game G for (5) with EDUs π_a, π_b and π_c; we'll compress the notation of Section 3.1 here and just represent 0's play ρ with the sequence $\pi_a.\pi_b.\pi_c$. For any discourse relation $\mathcal{R}$, we let $\neg\mathcal{R}$ denote the set of all relations $\mathcal{R}'$ such that $\mathcal{R}' \neq \mathcal{R}$. We have two players in the ME game, 0, the author, and 1, the reader. Player 0 first plays $\langle \pi_a \pi_b \rangle$, though she could have chosen a different play, say $\langle \pi_{a'} \pi_{b'} \rangle$, with different possibilities for discourse links between them. Now as 1 reads the article, he specifies the underspecified relations linking the EDUs of the text and hence builds up a history. For example, a plausible history might look like $\langle \pi_a, \pi_b \rangle \langle \mathsf{contr}(\pi_a, \pi_b) \rangle \langle \pi_c \rangle \langle \mathsf{res}(\pi_b, \pi_c) \rangle \ldots$ and so on. The various histories constitute the different branches of the resulting ME game structure abstractly depicted in Figure 3. As in equation (7), π below represents the CDU consisting of a and b.

We name these alternative ME game plays $\rho_1, \ldots, \rho_7$ as shown in the figure. ρ^0 will denote the empty history and for every $1 \leq i \leq 7,\ j \geq 1$, ρ_i^j will denote the prefix of the history ρ_i after round j of the ME game has been played. Thus, for example, $\rho_4^2 = \rho_5^2 = \langle \pi_a \pi_b \rangle \langle \neg \mathsf{contr}(a,b) \rangle$ (the two ME plays agree to this point).

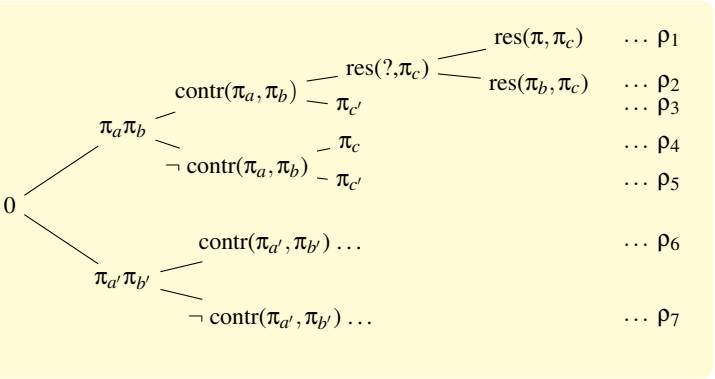

Fig. 3. A partial ME evaluation game tree for (5)

Let us now fix some player types. We envisage two types for 1, the 'left wing' and the 'right wing' type respectively denoted as t_1^ℓ and t_1^r. These are tied to winning conditions or sets of conversations, those conversations in which the march is characterized either as something interesting or worthwhile and those in which the march is characterized as something bad. We also assume that there are two relevant types for the author 0, the 'bad' type t_0^b, who is trying to convince the reader that the march for science is bad, and the 'good' type t_0^g, who is portraying the march as good. The strategies for 0 are to choose what aspects of the March to talk about and then to choose whether to connect them using relations signalled by the grammar or leave the connections underspecified for 1 to fill in. The strategies of 1 as he reads through the text connect every new discourse unit with the history he has built so far by assigning discourse relations, thus extending ρ when these connections are not explicit. In the context of the above simplified ME game, let's consider two strategies for 0: play $\langle\pi_a\pi_b\rangle$ or an alternative $\langle\pi_{a'}\pi_{b'}\rangle$, which could be the two opening clauses of (1). The strategies of 1, once 0 has played $\langle\pi_a\pi_b\rangle$ or $\langle\pi_{a'}\pi_{b'}\rangle$, are to relate them either with Contrast (contr) or something else. We denote the strategies as:

$$\sigma_1 \equiv \mathsf{contr}(\cdot,\cdot) \qquad\qquad \overline{\sigma}_1 \equiv \neg\mathsf{contr}(\cdot,\cdot)$$

Next, 0 plays π_c, which is among all the possibilities $\pi_{c'}$ which she could have played. Nevertheless, the two possible strategies for 1 after 0 has played π_c or $\pi_{c'}$ are to relate them with a Result (res) or some other relation. These are denoted as:

$$\sigma_2 \equiv \mathsf{res}(\cdot,\cdot) \qquad\qquad \overline{\sigma}_2 \equiv \neg\mathsf{res}(\cdot,\cdot)$$

Now suppose the prior beliefs of the two types of 1 are given as in the following table.

$\beta_1^\rho(\cdot,\rho^0)$	t_0^b	t_0^g
t_1^r	0.8	0.2
t_1^ℓ	0.5	0.5

Thus, the 'right wing' type t_1^r of the reader, who we'll assume is familiar with the conservative slant of *Townhall*, believes that the author is of the 'bad' type t_0^b with a higher probability of 0.8 whereas the 'left wing' type t_1^ℓ, who is not familiar with *Townhall*, is more neutral and yet undecided about the type of the author and the impact of the march for science in general. So they give the author the benefit of the doubt by assigning equal probabilities of them being either 'good' or 'bad'.

Belief dynamics in ME games and bias hardening With the basics of ME evaluation games in place, we now illustrate how the beliefs of our players evolve as the ME evaluation game progresses. Via the Types/History Correspondence, we can show a mutual hardening both of the interpreter's beliefs about types and his confidence in interpretations, producing *self-confirming* biases.[15]

To illustrate let us examine the interpreter type t_1^r reader and see how his interpretation of the plays, i.e., how the history which he builds as he reads the text, affects his initial beliefs. As we said above, t_1^r's beliefs assign an overall probability of .8 to 0's being of type t_0^b, and a probability of .2 to her being of type t_0^g. Now we see how that probability distributes over the various strategies, $\sigma_1, \overline{\sigma}_1, \sigma_2, \overline{\sigma}_2$, as defined above. As intuition suggests, we assume that t_1^r believes with certainty (probability 1) that the type t_0^b of 0 chose the play $\langle \pi_a \pi_b \rangle$ and intended the underspecified relation between π_a and π_b to be a Contrast with a rhetorical question as the first argument (contr). If we distribute the overall probability of .8 for the type t_0^b over the strategies $\sigma_1, \overline{\sigma}_1, \sigma_2, \overline{\sigma}_2$, this corresponds to 0.4 for σ_1 (and 0 for $\overline{\sigma}_1$) in the table below. The fact that t_1^r's belief that (b) and (c) are related by Result (strategy σ_2) is slightly less certain but still very strong, with a probability of .99, is also factored into the table below for t_0^b. The probability .2 for t_0^g is distributed equally over the two strategies $\overline{\sigma}_1$ and $\overline{\sigma}_2$: t_1^r believes with certainty that the type t_0^g of 0 would not have played $\langle \pi_a, \pi_b \rangle$ and thus intended the two underspecified relations as not being $\mathsf{contr}(\pi_a, \pi_b)$ or $\mathsf{res}(\pi_b, \pi_c)$. We can then sum up the prior interpretations of t_1^r of the play ρ (i.e., $\pi_a.\pi_b.\pi_c$) in the following table.

$\xi_1^\rho(t_1^r, \cdot)$	σ_1	$\overline{\sigma}_1$	σ_2	$\overline{\sigma}_2$
t_0^b	0.4	0	0.396	0.004
t_0^g	0	0.1	0	0.1

Now, let E_1^r be the event $\{\sigma_1\}$ and E_1^ℓ be the event $\{\overline{\sigma}_1\}$. We assume that the prior beliefs of the players are updated by Bayesian updates. Now, if we assume that event E_1^r occurs, and that E_1^ℓ can thus not occur, the prior beliefs of t_1^r get updated to yield:

$$\beta_1^\rho(t_1^r, \rho_1^2)(t_0^b) = \beta_1^\rho(t_1^r, \rho^0)(t_0^b | E_1^r) = 0.8/0.9 = 0.88$$

and

$$\beta_1^\rho(t_1^r, \rho_1^2)(t_0^g) = 1 - \beta_1^\rho(t_1^r, \rho_1^2)(t_0^b) = 0.12$$

[15] Self-reinforcing biases of nonlinguistic facts are also echoed in popular analyses, for instance 'The Evangelical Roots of Our Post-Truth Society' by Molly Worthen, *New York Times*, 16.04.2017. But as far as we know, only [8] has provided at least a partial formal analysis of this phenomenon.

So after $\langle \pi_a, \pi_b \rangle$ has been played and t_1^r has interpreted the play by constructing the history in ρ_1^2, t_1^r's updated beliefs and interpretations, noted with some abuse of notation as $\xi^{\rho_1^2}(t_1^r, .)$, (where $\xi_1^{\rho_1^2}(t_1^r, t_0^b)(\sigma_2) = .88 \times .99 = .871$ and where all of $\xi^{\rho_1^2}(t_1^r t_0^g)$'s updated probability mass is now on $\overline{\sigma_2}$) are given as in the following tables.

$\beta_1^\rho(\cdot, \rho_1^2)$	t_0^b	t_0^g
t_1^r	0.88	0.12

$\xi_1^{\rho_1^2}(t_1^r, \cdot)$	σ_2	$\overline{\sigma}_2$
t_0^b	0.871	.009
t_0^g	0	0.12

Similarly, let E_2^r be the event $\{\sigma_2\}$ and E_2^ℓ be the event $\{\overline{\sigma}_2\}$. When event E_2^r occurs, the prior beliefs of t_1^r get updated again as:

$$\beta_1^\rho(t_1^r, \rho_1^4)(t_0^b) = \beta_1^\rho(t_1^r, \rho_1^2)(t_0^b | E_2^r) = 0.87/0.88 = 0.988$$

and

$$\beta_1^\rho(t_1^r, \rho_1^4)(t_0^g) = 1 - \beta_1^\rho(t_1^r, \rho_1^4)(t_0^b) = 0.012$$

Our simple analysis here examines just the first 3 EDUs, and without many options for continuations, t_1^r's beliefs about 0's type quickly converge to probability 1. But in principle, we could carry out this analysis for the entire text or even a potentially infinite discourse. If the type of player 1 assigns a non-zero probability to the strategy that they do not believe their opponent will play and if they keep updating their beliefs based on Bayesian updates, we can show that their bias will get more and more skewed and eventually reach probability 1. This can be pictorially shown as in Figure 4. The intuition is that player 1 started off the game with strongly biased priors and these priors lead him to interpret the plays (construct histories) as he sees fitting with his initial bias, which then skews his beliefs even further.

Strategic consequences of bias hardening Given our analysis of bias hardening, player 1's reinforced belief that 0 is of type t_0^b should be a predictable outcome of 0's play, her choice of events to describe in example (5) and how she chooses to describe them. Why would 0 choose such a strategy? It stands to reason that there is a strong statistical correlation between t_0^b and the right wing t_1^r, something we call *type compatibility*. Type compatibility is reflected in t_1^r's priors for t_0^b. Let e_b and e_g stand for the March's being a bad thing or a good thing respectively, and let $\beta_1^*(t_1^r, .)$ extend $\beta_1^\rho(t_1^r)$ to types for events and states described in the plays or implied by them.

Assuming that t_0^b sincerely expresses her beliefs, $\beta_1^*(t_0^b, \varepsilon)(e_b)$ should be high (we'll assume it's 1); though t_0^b and e_b are different types about different individuals they convey the same message about e. We say two types are type compatible if their distributions are roughly aligned, and we say that t_1^r is type compatible with t_0^b. Thus, the prior $\beta_1^*(t_1^r, \varepsilon)(e_b)$ will already be high. As 1 updates his beliefs relative to a history that he has constructed for (5), the updated $\beta_1^*(t_1^r, \mathfrak{h}(\rho_1^4))(e_b)$ should be even higher, as 0's play supports and reinforces what 1 already believes. In response to the full $\mathfrak{h}(\rho_1^4)$, 1 has two strategies: σ_y (play "Like") and σ_n (play "Nay"). Given the type compability link we've just described, it is natural to conclude that $\beta_1^\rho(t_1^r, \mathfrak{h}(\rho_1^4))(\sigma_y)$ will go to 1, as his belief in e_b and in t_0^b go to 1. 0 is assured of a win by the Jury in such a game.

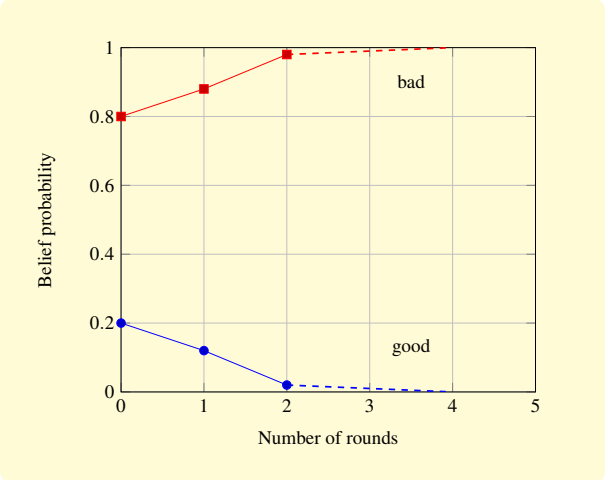

Fig. 4. The progressive reinforcement of bias

There are, however, other paths to victory for 0 if her partner is t_1^r in an ME evaluation game. Supposing $1's$ prior is such that $\beta_1^*(t_1^r)(e_b) = \beta_1^*(t_1^r)(e_g)$, the conditional probability of $\mathfrak{h}(\rho_1^4)$ given e_b is most likely close to 1, and this will increase the marginal probability of e_b, if t_1^r takes the source *Townhall* to be credible. In that case Bayesian updating will reinforce t_1^r's belief in e_b to the detriment of e_g, as we would expect, and repeating the last steps of reasoning above, 0 is once again assured of a win.

1's history ending with a "Like" should confirm 0's beliefs in t_1^r over t_1^ℓ. Whereas for instance t_0^b might have had relatively balanced priors over t_1^r and t_1^ℓ, $\beta_0^\rho(t_0^b, \mathfrak{h}(\rho_1^5))(t_1^r)$ should be significantly higher than $\beta_0^\rho(t_0^b, \mathfrak{h}(\rho_1^5))(t_1^\ell)$. And if 1 returns a "Like" after each turn by 0 followed by 1's history for the play at that point, the dynamics on 0's distribution over t_1^r over t_1^ℓ will pattern just like t_1^r's beliefs over t_0^b and t_0^g—namely, $\beta_0^\rho(t_0^b, \mathfrak{h}(\rho_1^n))(t_1^r)$ will converge to 1 as n increases.

Now let's allow β^*, the extended belief function, to provide a distribution over a subset of t_1 types when it takes as an argument a certain "mixed" type for 1, t_1^m. Intuitively t_1^m represents the fact that 1 may be of several minds about the story. Suppose now that t_1^m has a relatively balanced distribution over t_1^r and t_1^ℓ as in the paragraph above. Now if 1 behaves in the same way returning a "Like" after each turn by 0 by the same argument as above, t_1^m will converge to putting all the probability mass for t_1 types on t_1^r: $(\beta^*)_1^\rho(t_1^m, \mathfrak{h}(\rho_1^n))(t_1^r)$ will converge to 1 as n increases!

Observation 1 *In an ME evaluation game, 1's own biases over different types he may instantiate can shift and harden.*

We can thus model the idea that a text like (2) can "harden" the bias of 1—pushing him more towards the beliefs of t_1^r from another type. Moreover, the ME evaluation game framework shows that bias hardening could influence the winning condition for authors

of a continually updated body of testimony such as that provided by social media sites like *Facebook*. [4] show that such hardening makes the interpreter ever more reliant on that body of testimony.

Let's now look at the situations where 1 has type "left", t_1^ℓ. If we assume that like the type t_1^r, t_1^ℓ will interpret $\langle \pi_a, \pi_b \rangle$ as linked by Contrast, this interpretation will shift β_1^ρ to assign a higher value to the type t_0^b and accordingly to the type t_0^r. An interpreter of type t_1^ℓ may very well construct the same history $\mathfrak{h}(\rho_1^4)$ for the play in example (5) and come to the same reinforced belief in t_0^b as would an interpreter of type t_1^r, although more slowly. But this can work now to 0's disadvantage. Player 1 of type t_1^ℓ may wonder about the choice of events 0 relates; he may wonder what the scientists are saying about the March. More precisely, t_1^ℓ's type could be such that as a belief in t_0^b becomes reinforced via the interpretation of (2), t_1^ℓ will require more and more evidence that the history unfolding is in fact an accurate portrayal of the March. As (2) does not argue its case, the more the interpretation diverges from an evenhanded or argued evaluation, the more t_1^ℓ will not be persuaded. In such a situation, 0 has not gauged her opponent well; she didn't take into account t_1^ℓ in her strategy, as her play leads to a suboptimal result. Recall that 1 has two strategies at the end of the story and the history constructed from it: σ_y (play "Like") and σ_n (play "Nay"). Given our description of t_1^ℓ, $\xi_1^\rho(t_1^\ell, t_0^b)(\sigma_n)$ becomes stronger and stronger as 1 gathers evidence from reading (2). At the end of (2) $\xi_1^\rho(t_1^\ell, t_0^b)(\sigma_n) >> \xi_1^\rho(t_1^\ell, \mathfrak{h}(\rho_1^n)(\sigma_y)$; 1 says *Nay*, and the Jury accords 0 a loss.

There are still other ways this interpretation and belief shift could go. But we have demonstrated that 0's success in an ME evaluation game depends on her assessing correctly player 1's type. We could complicate our system of types and distinguish two right wing types for 1, the t_1^r type that has right wing views and holds them to be self-evident and a more moderate right wing type t_1^{rm} that may also suspect that the March for Science is bad but for whom the position requires a reasoned argument to be persuasive. Player 0 as t_0^b has a winning strategy against t_1^{rm}, with a play more like example (3). The space of ME evaluation games reflects probabilities of achieving a winning condition in measure theoretic terms; if 0 has a probability of winning an ME game, then she almost always has a winning strategy. Conversely if 1 has a probability of saying "Nay", then 0 almost always loses the game. Summing up our findings:

Observation 2 *1. If* 1 *can infer that* 0 *is type compatible with* 1*'s type in an ME evaluation game* $\mathcal{G}$*, then* 0 *almost always has a winning strategy in* $\mathcal{G}$.

2. If 0*'s beliefs concerning* 1*'s type are accurate in an ME evaluation game* $\mathcal{G}$*,* 0 *almost always has a winning strategy in* $\mathcal{G}$.

To construct a history from ambiguous signals in a play over X, the interpreter must rely on her beliefs about the situation and her interlocutors to estimate the right history. So the question of getting the histories the author intended depends at least in part on the right answer to the question, what are the right beliefs about the situation and the participants that should be invoked in interpretation? Given that beliefs are probabilistic, the space of possible beliefs is vast. The right set of beliefs will typically form a very small set with respect to the set of all possible beliefs about a typical conversational setting. Assuming that one will be in such a position "by default" without any further argumentation is highly implausible, as a simple measure theoretic argument ensures that the set of possible interpretations are almost always biased away from capturing a

particular intended interpretation in an ME truth game. An immediate consequence of this is that misinterpretations, where the interpreter constructs a history from the play in an ME game that is different from the history the author intended are expected.

Of course, this is known to the authors of our histories. Player 0 in an ME game can intentionally send an ambiguous message via a play ρ without intending to communicate a particular history, allowing the interpreter, player 1, to take what he wishes from ρ. But, provided player 0 wants to convey a particular history $\mathfrak{h}_0$ via ρ, the discourse connections between discourse units and the specifications of other underspecified elements in ρ must be reliably recoverable for 1, in order for him to reproduce $\mathfrak{h}_0$. This can happen in two ways: player 0 knows that 1 has detailed knowledge of her type (which means *inter alia* that 1 knows what type 0 assigns to 1). Alternatively, ρ must exploit linguistic devices to make the recovery of $\mathfrak{h}_0$ largely a matter of decoding linguistic signals. As a result, ρ will reveal relevant information about her type through her play, in virtue of the type/history correspondence. Thus,

Observation 3 *In an ME evaluation game $\mathcal{G}$ if* 0 *intends to convey $\mathfrak{h}_0 \in \mathfrak{h}(\rho)$ for a general audience (without specific knowledge of the type of player 1), then $\{\mathfrak{h}_0\} = \mathfrak{h}(\rho)$—i.e., $\mathfrak{h}_0$ must be largely grammatically or linguistically determined.*

And as a history includes discourse structure, 0 will, given our assumptions, make most of the discourse connections in a history explicit, as in our stories above,

Corollary 1 *The discourse structure of a history for a general audience in an ME evaluation game, where* 0 *intends to convey a particular history, must be largely grammatically or linguistically determined.*

ME games predict more explicit connections in a piece for a general audience as in our newspaper stories than, say, in a conversation between two people who know each other well. That is in evidence in our analysis of (2); although we showed that beliefs about types evolved as interpretation progressed, the interpretation of the text was pretty much grammatically fixed. Quantitative evidence for the predictions in Observation 3 and its Corollary 1 comes from a comparison of the number of discourse connectors, explicit indicators of discourse relations, in corpora of newspaper texts like the RST Tree Bank and corpora of chat conversations like the STAC corpus; chats in general have far fewer explicitly marked discourse connections, and rely more on information about the situation and about conversational participant types. However, there is a way that histories though grammatically determined can have their content affected by beliefs about types. We study this in the next section.

4.2 Epistemic content

In an ME evaluation game, player 0 may choose to leave certain elements underspecified or ambiguous, or use a specified construction, to invoke epistemic content for a particular type τ that she is confident player 1, the interpreter, instantiates. How much so depends on her confidence in player 1's type being τ—on the probability mass she accords to τ in her probability distribution over types. This confidence level opens a panoply of options about the uses of epistemic content: at one end there are histories

constructed from linguistic cues with standard, grammatically encoded meanings; at the other end there are histories generated by a code shared with only a few people whose types are mutually known. As the conversation proceeds, our ME evaluation games have shown that probabilities about types are updated and so the model predicts that a speaker can resort to more code-like messages in the face of feedback confirming her hypotheses about player 1's type (if such feedback can be given using what 1 has at her disposal) and that the speaker may revert to a message exploiting grammatical cues in the face of feedback disconfirming her hypotheses about player 1's type. Thus, the epistemic ME model predicts a possible change in register as the speaker receives more information about player 1's (or eventually the Jury's) type, though this change is subject to other conversational goals coded in the speaker's victory condition for the ME game.

A limiting case of this is what linguists and others call a *dog whistle*. In a dog whistle the author relies on the interpreter's type to add a meaning over and above an already fixed, literal meaning of what she says. Consider the following move by Marion Le Pen, a leader of the French nationalist, right-wing party *le Front National* in which she recently said:

(10) La France était la fille aînée de l'église. Elle est en passe de devenir la petite nièce de l'Islam. (*France was once the eldest daughter of the Catholic church. It is now becoming the grand niece of Islam.*)[16]

(10) sets up a contrast between what once was, on the one hand, and what will soon be, on the other in terms of faiths in France using a metaphor that on the face of it is a little strange but could be interpreted innocently as a description of the increasing numbers of French practioners of Islam and the move to a religiously more pluralistic society. But this is not an innocent contrast for a certain type. (10) exploits her intended audience's type and the associated beliefs about Islam, Catholicism and France. She assumes that her interpreters' type will know about the Pope John Paul II's characterization of France as the eldest daughter of the Church, hearkening back to a phrase used in medieval and Renaissance times. The phrase will come loaded with many historical associations to a time when France was a pre-eminent power. Le Pen then contrasts that time with the current and near future where France has a subservient position as a grand niece of Islam, which the intended type will associate with people of North African descent. With these background beliefs, Le Pen's statement takes on an additional, loaded racist meaning, conveying an assault on France and its once proud status by people of North African descent. Thus by exploiting the interpreter's type, the author of such a discourse move communicates a much more determinate content than what is grammatically determined [31].[17]

---

[16] https://www.youtube.com/watch?v=9r8fKymWWZ8

[17] While [56] proposes that such messages are conventional implicatures, [31, 36] show that dog whistle content doesn't behave like other conventional implicatures; in terms of tests about "at issue content", dog whistle content patterns with other at issue content, not with the content associated with conventional implicatures in the sense of [52].

Dog whistles add content, a loaded interpretation, to a specific discourse unit that goes beyond its grammatically determined meaning.[18] This particular content arises because of the interpreter's or hearer's type. In line with [21], the use of the historical expression *la fille ainée de l'église* contrasted with *la petite nièce* in (10) comes to encode a type about the interpreter; 1 will believe that 0 has the strategy of using just this language to make the loaded interpretation accessible and moreover will identify with its content.

Our framework is ready made to encode these observations formally. Recall that a play ρ in an ME game may be compatible with several fully specified histories given an interpreter's type t_1. Let $\mathcal{H}(\rho)_{t_1}$ be the set of histories compatible with a play ρ given an interpreter type t_1.

Definition 11. ρ *will be ambiguous and open to epistemic content supplementation just in case: (i)* $|\mathcal{H}(\rho)|_{t_1} > 1$ *for any type* t_1 *for a linguistically competent interpreter, and (ii) there are* $\mathfrak{h}_1, \mathfrak{h}_2 \in \mathcal{H}(\rho)_{t_1}$*, such that* $\mathfrak{h}_1$ *and* $\mathfrak{h}_2$ *are semantically distinct (neither one entails the other).*

A dog whistle play ρ in its loaded interpretation yields a separate history $\mathfrak{h}^{\dagger}_{\rho}$ from the history $\mathfrak{h}^{*}_{\rho}$ that encodes the "unloaded" or literal interpretation. However, a dogwhistle is not ambiguous in the standard sense. The additional meaning is typically related to the literal meaning via a discourse relation, the relation being needed to ensure coherence of the content as a whole. Consider our example (10); the additional racist meaning comes as a commentary or explanation on the more neutral meaning: France was once the eldest daughter of the Catholic church. It is now becoming the grand niece of Islam *because Northern Africans are overrunning France*. Or consider an alleged dogwhistle like the one discussed in [15] in which the term *welfare*, whose definition might be *government support intended to ensure that members of a society can meet basic human needs*, conveys the alleged dogwhistle content *Black people are lazy*. The full content of the dogwhistle interpretation actually involves the literal content and the supplementary content in a mini discourse structure; e.g., whenever the term welfare is used it evokes a meaning something like *government support intended to ensure that members of a society can meet basic human needs, which is needed because Blacks are lazy*. The relative clause provides a commentary on the notion of welfare.

Definition 12. ρ *is a* dog whistle *for* t_1 *just in case: (i)* $|\mathcal{H}(\rho)|_{t_1} > 1$*, (ii)* $\mathfrak{h}^{*}_{\rho} \in \mathcal{H}(\rho)$ *and (iii) there is a* $\mathfrak{h}^{\dagger}_{\rho} \in \mathcal{H}(\rho)_{t_1}$ *with epistemic content* ϕ *that can positively affect some jury perhaps distinct from* t_1 *and such that for some veridical discourse relation R, some* π *occuring in* $\mathfrak{h}^{*}_{\rho}$*,* $\mathfrak{h}^{\dagger}_{\rho}$ *contains* $\mathfrak{h}^{*}_{\rho}$ *and the formula* $R(\pi, \pi') \wedge \langle \pi' : \phi \rangle$.

Note first that the interpreter, player 1, will acquire the dogwhistle type only if he is of a certain type t^{d}_{1}, a type that has the requisite beliefs needed to build $\mathfrak{h}^{\dagger}_{\rho}$. Note also that in a dogwhistle play ρ, $\mathfrak{h}^{\dagger}_{\rho} \models \mathfrak{h}^{*}_{\rho}$, since for R to be veridical, $R(\pi, \pi') \wedge \langle \pi' : \psi \rangle \models \psi$, where $\models$ is the semantic consequence relation defined in SDRT and hence applicable to histories and plays. According to our definition, dogwhistles can be unintentional [53],

[18] This in fact is a precise model of what relevance theorists have called "free enrichment" [55].

as the conditions under which they are triggered has to do with the interpreter's beliefs not with the author's, but their interesting strategic use is as intentional, rhetorical devices.

Suppose a type for 0, call it t_0^d, believes that in an ME evaluation game G, $Win_0 = \{x.\mathfrak{h}_\rho^\dagger.\{\langle like\rangle\}\}$, with $x \in (V_0 \cup V_1)^*$, which entails that 1 must be of type t_1^d and in addition must have the property that at the end of the play ρ, $\xi_1^\rho(t_1^d,t_0^d)(\langle like\rangle)$ will be very high. An immediate consequence of this is:

Observation 4 *In an ME evaluation game $\mathcal{G}$, if Win_0 in $\mathcal{G}$ consists in conveying $\mathfrak{h}^\dagger \in \mathcal{H}(\rho)_{t_1}$ that is strictly stronger semantically than the grammatically determined $\mathfrak{h}_\rho^* \in \mathcal{H}(\rho)_{t_1}$, then 0 has a winning strategy in $\mathcal{G}$ only if she knows 1's type is t_1^d.*

Observation 4 shows that dog whistle epistemic content, to be reliably conveyed, requires a knowledge of the interpreter's type. On the other hand, if 0 utters a dog whistle play ρ, she has a defense when challenged about its content: she can say she actually meant $\mathfrak{h}_\rho^*$—which is an essential feature of a dog whistle. The notion of an ME persuasion game that we define in the next section will make this clear.[19]

We've seen that epistemic content can fill in grammatically underspecified meanings, and we've seen it can also lay an additional layer of meaning on top of grammatically specified meaning. The last type of epistemic content we discuss has to do with the *semantic bleaching* of grammatically determined content, a process whereby the interpreter "erases" literal content. Semantic bleaching is thus a kind of inverse of a dog whistle. Consider an ME evaluation game with example (1) as a play and where 1 is of type t_1^r. If 1 infers that the author of (1) is of type t_0^ℓ and hence of incompatible type, a natural reaction is to eventually stop attending to the literal meaning of (1), explaining his interpretive behavior with a phrase like *well, it's all just left wing cant*. In so doing 1 assigns the text a vacuous content. Similarly our left wing interpreter of (2) may dismiss that story, according it a similarly vacuous content. As such 0 will clearly lose in the corresponding ME evaluation game.

Definition 13. *In an ME game, player i's play or element of a play ρ is* semantically bleached *for player or Jury type k iff for the most plausible histories for k, $\mathfrak{h}_k(\rho)$—i.e. for all alternative histories $\mathfrak{h}_a(\rho)$ such that $\xi_k^\rho(T_k)(\mathfrak{h}_k(\rho)) > \xi_k^\rho(T_k)(\mathfrak{h}_a(\rho))$— ρ introduces a discourse unit π into ξ_k^ρ and the V_i formula $(\langle\pi\colon \top\rangle, i)$, with $\top$ signifying a discourse move with vacuous content.*

4.3 ME persuasion games

The ME evaluation games we have used up to now in exploring interpretive bias are relatively simple. They are very "one-sided" conversations, since 1's options for expressing himself are limited to "like" and "nay". They have served, however, to see how biases in the form of probability distributions over types can affect interpretation and how those can lead to self-reinforcing biases for the interpreter. ME evaluation games have also served to model the strategic consequences of author and interpreter biases and to elucidate dog whistles and other forms of epistemic content.

[19] As far as we know, no other framework can capture the observations below.

ME games also allow us to model more complex strategic interactions. We now explore a more sophisticated ME game that builds on the idea of an ME evaluation game. Like all the ME games we consider here, it is distinguished from other types of ME games by the winning condition. In an ME persuasion game, 0's goal is to persuade the Jury to accept her history, while 1's objective is to poke holes in 0's history.

Definition 14. *An ME persuasion game is an epistemic ME game with two players,* 0 *and* 1*, in which* 0 *furnishes a play* ρ *that develops and defends a history* $\mathfrak{h}_0$ *or set of histories in* $\mathcal{H}(\rho)$ *against attacks and questions by* 1*.* 0 *wins iff the Jury decides that* 0 *has defended some* $\mathfrak{h} \in \mathcal{H}(\rho)$ *adequately against attacks from* 1*;* 1 *wins iff* 0 *loses.*

In an ME persuasion game, 0 makes a play ρ in which she tries to defend a history or set of histories $\mathfrak{h}_0$ in $\mathcal{H}(\rho)$, and 1 tries to show $\mathfrak{h}_0$ is incorrect, misguided, based on prejudice or whatever he thinks will convince the Jury to be dissuaded from adopting 0's view of some body of data X. 1 may present elements of X that 0 has left out, as well; he may fill in missing elements to what he perceives to be his advantage, although 0 is also free to object to specifications of 1 that she does not wish to defend. In so doing, she in turn fills in the underspecified elements of her play.

As in all ME games, ME persuasion games are determined by player victory conditions given by the Jury; we define them above as win/lose games, although we could specify victory conditions in which both participants fail to achieve their winning conditions. ME persuasion games are more complex than ME evaluation games, however, for several reasons. First, whereas player 1 serves as the Jury in an ME evaluation game, the Jury in an ME persuasion game is an external evaluator that serves only to determine the winning conditions of the game and thus pick the winner in 0's and 1's debate. Second, while in ME evaluation games, player 0 receives little information from player 1, in ME persuasion games, player 1 can offer detailed counterarguments, which affect both his, 0's and the Jury's beliefs, their interpretations and their strategies.

In addition, in ME persuasion games, unlike ME evaluation games, the Jury's type can vary independently of player 1's type. Thus, 0's and 1's strategies crucially depend on their beliefs about the Jury: 0 has to construct $\mathfrak{h}_0$ in ways that will please the type she believes the Jury has; 1 has to attack $\mathfrak{h}_0$ in ways that accord with his beliefs about the Jury; and 0 has to defend $\mathfrak{h}_0$ in ways that will, given her beliefs, dispose the Jury favorably to it. In particular, different Jury types can differ over what constitutes a winning history over some common set of data X. A Jury of type t_J^r would find the history $\mathfrak{h}$ suggested by 0 with the play in example (2) acceptable, *despite whatever arguments player* 1 *might advance against* $\mathfrak{h}$. A Jury of type t_J^ℓ might have the reverse judgment and be all too ready to accept any argument player 1 might advance against $\mathfrak{h}$. In contrast to Aumann's dictum [11], there is every reason for these two Juries to agree to disagree![20] If the Jury's type is t_J^r, for instance, then the arguments about bias hardening apply to the Jury and we have the correlates of Observations 1 and 2 for ME persuasion games:

Observation 5 *1. If the Jury can infer that 0 is of a compatible type in an ME persuasion game G, then 0 almost always has a winning strategy in G.*

[20] Technically, Aumann's observation relies on common prior probabilities. We don't see any reason to adopt such an assumption in an analysis of strategic conversations or bias. Our observation is a sort of correlate or converse of Aumann's.

2. *If* 0*'s beliefs concerning the Jury's type* $t_{\mathcal{J}}$ *are accurate in an ME persuasion game* $\mathcal{G}$*, then* 0 *has a high probability of having a winning strategy in* $\mathcal{G}$*.*

If we impose certain constraints on the Jury type, then 1 can have an effect on 0's winning condition. For instance let us suppose that the Jury is of a type that rejects any history/play iff it is demonstrated to have some inadmissable property P (racist, sexist, ...). Call such a jury type the *enlightened Jury type* t^e_j.

Observation 6 *Suppose an ME persuasion game* $\mathcal{G}$ *with the Jury of type* t^e_j *in which* 1 *can show* $\mathfrak{h} \models P(\mathfrak{h})$*, with* P *an inadmissable property and with* $\mathfrak{h}$ *the history* 0 *has proposed. Then* 0 *has no winning strategy in* $\mathcal{G}$*.*

Let's now turn to see how different constraints on Jury type will make different sorts of epistemic content have a strategic use in ME persuasion games. The use of a dog whistle can plainly affect a like minded Jury positively without saddling 0 with the provocative interpretation of the dog whistle. Suppose an ME persuasion game $\mathcal{G}$ in which the Jury is of type t^e_j and 0 can use a play ρ for which there are two histories $\mathfrak{h}^*_\rho, \mathfrak{h}^\dagger_\rho$, the first literal, the second supplemented with some inadmissable epistemic content. 0 can respect the Jury's sense of "fair play" while still conveying the objectionable content to a like minded audience. In the case of dog whistle content, it becomes more difficult for 1 to prove that 0's proposed history has an inadmissable property solely in virtue of the dog whistle content. 0 can deter attacks by 1 of impropriety, racism, sexism and so on, because 0 can say that the dog whistle content is not something she is responsible for but comes from epistemic supplementation.

Observation 7 *Suppose an ME persuasion game* $\mathcal{G}$ *with a Jury of type* t^e_j *and a dog-whistle play* ρ *such that there are* $\mathfrak{h}^*_\rho, \mathfrak{h}^\dagger_\rho \in \mathcal{H}(\rho)$ *with* $\mathfrak{h}^\dagger_\rho \models \mathfrak{h}^*_\rho$ *and for some inadmissable property* P*, player* 1 *can show that* $\mathfrak{h}^\dagger_\rho \models P(\mathfrak{h}^\dagger_\rho)$*, but in fact* $\mathfrak{h}^*_\rho \not\models P(\mathfrak{h}^*_\rho)$*. Then* 0 *has a winning strategy in* $\mathcal{G}$*.*

0's winning strategy consists in convincing the Jury that $\mathfrak{h}^*_\rho \in \mathcal{H}(\rho)$ and $\mathfrak{h}^*_\rho \not\models P(\mathfrak{h}^*_\rho)$ for any inadmissable property P that 1 proposes.

On the other hand, semantic bleaching can negate all of 0's efforts as the consideration of the following ME persuasion game shows. Suppose an ME persuasion game $\mathcal{G}$ where the play is (1) and 0 has type t^g_0, and say that a jury $\mathcal{J}$ and player i are *type incompatible* just in case given some ρ, $\xi^\rho_{\mathcal{J}}(t_1,t_0)(\langle like\rangle) = 1 - \xi^\rho_i(t_1,t_0)(\langle like\rangle)$; that is, they assign inconsistent interpretations of ρ. Let the Jury be of type $t^s_{\mathcal{J}}$, where t^s_j always induces semantic bleaching of contributions by any player i with whom it is type incompatible, and assume that t^s_j has the same priors as the right wing type t^r_1 from the previous section. It seems reasonable to conclude that t^g_0 and t^r_j are type incompatible, and thus that no amount of explanation or rebuttal by 0 to an attack by 1 will lead to a winning strategy for 0 in $\mathcal{G}$. 0's efforts will simply contribute nothing to the conversation because of semantic bleaching, and the Jury will accord 1 the win.

Observation 8 *Suppose an ME persuasion game G where the Jury is of type* t^s_j *and type incompatible with* t_0*. Then* 0 *has no winning strategy in G.*

As a final illustration of the importance of the Jury's type in an ME persuasion game, the Jury's type can specify what it means for 0 to have mounted an adequate defense against attacks by 1. For instance, the Jury's type could specify that the proportions of good unanswered attacks in an ME persuasion game $\mathcal{G}$ on the latest version of 0's play with respect to the total number of attacks at some point continues to diminish and eventually goes to 0. At each turn n of the game, 1 can argue about the history prefix or prefixes $\mathfrak{h}_n$ constructed by 0 so far, challenge them with new facts or attack its assumptions (for a definition see [8]), with the result that 0 may rethink and redo portions of $\mathfrak{h}_n$ (though not abandon the original history entirely) in order to render 1's attack moot. This winning condition is a sort of limit condition: if we think of the n initial segments of 0's play as producing an "initial" play $\mathfrak{h}^0_\rho$ over $\mathcal{X}$, as $n \to \infty$, ρ^0_n has no unanswered counterattacks by 1 that affect the Jury.

We can then characterize the complexity of this winning condition using the mathematical structure of an ME game. More precisely, where a *good attack* by 1 is one to which 0 does not have a convincing response for the Jury and ρ_n is a prefix of play ρ of length n, 0's winning condition is then specified by:

$$(11) \qquad Win_0 \;=\; \{p \in (V_0 \cup V_1)^\infty \mid \limsup_{n\to\infty} \frac{\text{good attacks by } 1 \text{ in } \rho^0_n}{\text{attacks by } 1 \text{ in } \rho^0_n} = 0\}$$

This precisification makes 0's winning condition extremely complex.[21] As one can see from inspection, no finite segment of an infinite play guarantees such a winning condition. We shall call an initial segment of a history in 0's winning condition as we have just characterized it, 0-*defensible*. A weaker condition on 0's winning condition imposed by the Jury might simply be that 0's play be 0-defensible to a certain length (known perhaps only to the Jury).

ME persuasion games invite a study of attacks, which can draw on work in argumentation and game theory [24, 29, 18]. ME persuasion games go beyond the work just cited, however, in several ways: attacks in ME games involve complex linguistic behavior and conversational interactions; the notion of an effective attack, as we have seen, involves the type of the Jury as a crucial parameter; and the effectiveness of an attack for a Jury relies on its prejudices. This last point is obvious once one thinks about it, but to our knowledge this feature of ME games is not a part of argumentation theory.

4.4 ME truth games

We now turn to a special kind of ME persuasion game, an ME *truth game*. An ME truth game will allow us to define what an "optimal bias" might look like, where we understand an optimal bias as one that gets as close as possible to ground truth. One key notion of an ME truth game is a *disinterested* Jury. The intuition behind a disinterested Jury is simple: such a Jury judges the persuasion game based only on the public commitments that follow from the discourse moves that the players make; it is not predisposed to either player in the game. While it is difficult to define such a disinterested Jury in terms of its credences, its probability distribution over types, we can

[21] Using the mathematical structure of ME games in [9], in which winning conditions can be characterized in terms of the Borel hierarchy, we can show that this winning condition is at least in the Π^0_3 level of the Borel Hierarchy and not first order definable [23].

establish some necessary conditions. We first define the notion of the dual of a play of an ME game. Let $(v,i) \in (V_0 \cup V_1)$ be an element of the labeled vocabulary with player $i \in \{0,1\}$. Define its dual as:

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

The dual of a play $\rho \in (V_0 \cup V_1)^\infty$ 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 $x_0 = \varepsilon$ then

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

Then, a disinterested Jury must necessarily satisfy:

- **Indifference towards player identity:** A Jury $\mathcal{J} = (Win_0, Win_1)$ is unbiased only if for every $\rho \in (V_0 \cup V_1)^\infty$, $\rho \in Win_i$ iff $\overline{\rho} \in Win_{(1-i)}$.
- **Symmetry of prior belief:** A Jury is unbiased only if it has symmetrical prior beliefs about the player types.

While symmetry of prior beliefs is satisfied by a uniform distribution over all types, it does not entail such a uniform distribution; it requires that the prior distribution be the same or approximately the same for both players.[22] Symmetry is closely related to the principle of maximum entropy used in fields as diverse as physics and computational linguistics [17], according to which the absence of any information about the players would entail a uniform probability distribution over types.

The notion of a disinterested jury is formally a complicated one. Consider an interpretation of a conversation between two players 0 and 1. Bias can be understood as a sort of modal operator over an agent's first order and higher order beliefs. So a disinterested Jury in an ME game means that neither its beliefs about 0 nor its beliefs about 1 involve an interested bias; nor do its beliefs about 0's beliefs about 1's beliefs or 0's beliefs about 1's beliefs about 0's beliefs, and so on up the epistemic hierarchy. Thus, a disinterested Jury in this setting involves an infinitary conjunction of modal statements, which is intuitively (and mathematically) a complex condition on beliefs. And since this disinterestedness must be common knowledge amongst the players, 0 and 1 have equally complex beliefs.

A disinterested Jury should evaluate a conversation based solely on the strength of the points put forth by the participants, and crucially, it should evaluate the conversation in light of the *right* points. Appeals to ad hominem attacks by 1 or colorful insults, for instance, should not sway the Jury in favor of 1; it should evaluate only based on how the points brought forward affect its credences under conditionalization. A distinterested Jury is impressed only by attacks from 1 that are based on evidence (0's claims aren't supported by the facts) and on formal properties of coherence, consistency and explanatory or predictive power. In such a game, it is common knowledge that attacks based on information about 0's type that are not relevant either to the evidential support or formal properties of her history are ignored by the Jury. The same goes for 0; counterattacks by her on 1 that are not based on evidence or the formal properties mentioned above should be ignored.

[22] For a fuller discussion of symmetry, see [8].

[9] discusses the formal properties of coherence and consistency in detail, and we say more about explanatory and predictive power below. But is the existence of an ME persuasion game G in which 0 elaborates and successfully defends a coherent and consistent history $\mathfrak{h}_{0,X}$ over a set of facts X with a disinterested Jury sufficient to a maximal approximation to the truth? No. We need to specify the evidential criteria the Jury uses. In addition, there may be a fatal flaw in $\mathfrak{h}_{0,X}$ that 1 does not uncover and that the Jury does not see. We must suppose certain abilities on the part of 1 and/or the Jury: if 0 has covered up some evidence or falsely constructed evidence or has introduced an inconsistency in $\mathfrak{h}_{0,X}$, that eventually 1 will uncover it; if there is an unexplained leap, an unanswered question, an incoherence in the history, then 1 will eventually find it.

Formal epistemologists have formulated constraints like *cognitive skill* and *safety* or *anti-luck* on beliefs that are relevant to characterizing this evidential criterion [48, 37]. Cognitive skill determines the success or accuracy of an agent's beliefs: it is the extent to which the reasoning process that produces the beliefs prioritizes evidential factors (how weighty, specific, misleading, etc., the agent's evidence is) and makes non-evidential factors comparatively unimportant. *Safety* or *anti-luck* is a feature of beliefs that says that conditionalizing on circumstances that could have been otherwise without one's evidence changing should not affect the strength of one's beliefs. Safety rules out belief profiles in which luck or mere hunches play a role. In addition, we will require that the relevant evidential factors are those that have been demonstrated to be effective in the relevant areas of inquiry.

If a Jury measures the success of a persuasion game in virtue of cognitive skill on the part of the participants and this is common knowledge among the participants (something we will assume throughout here), then, for instance, 1's attacks have to be about the particular evidence adduced to support 0's history, about the way it was collected or about verifiable errors in measurements etc. This requirement precludes general skeptical claims by 1 from being credible attacks. These epistemic components thus engender more relevant types for interpretation: are the players using cognitive skill and anti-luck conditions or not? For example, most climate skeptics' attacks on climate change science, appealing to general doubts about the evidence without introducing any credible scientific criteria attacking specific evidential bases, would be ruled as irrelevant in virtue of a property like cognitive skill. Such a criterion may also affect the Jury's interpretation of the conversation. A Jury whose beliefs are constrained by cognitive ability will adjust its beliefs about player types and about interpretation only in light of relevant evidential factors.

A player 1 who is both cognitively skilled and safe will have the requisite forensic properties we alluded to above. Cognitive skill and safety determine particular Bayesian belief updates; and as they have to do with beliefs and how these beliefs respond to plays in an ME game, they can be coded in the type system. Assigning these properties to the Jury as well will give us the constraints we need on an ME persuasion game to get at a best approximation of ground truth.

Definition 15. *An* ME truth game $\mathcal{G}$ *is an ME persuasion game with types for the Jury and players i that imply that the Jury is disinterested, cognitively skilled and anti-luck adept and that players i are cognitively skilled and anti-luck adept. The victory condition is as specified in equation (11).*

In an ME truth game G, if 0 wins by defending $\mathfrak{h}_{0,X}$ over a set of facts X, then $\mathfrak{h}_{0,X}$ is as close epistemically as we can get to ground truth about X. The probability that $\mathfrak{h}_{0,X}$ is true given that 0 has won an ME truth game is high for any cognitively skilled player, high enough for rational acceptance of $\mathfrak{h}_{0,X}$. If 0 doesn't win, $\mathfrak{h}_{0,X}$ may still be the ground truth, but 0 cannot defend it as such.

Proposition 1. *In an ME truth game G, the Jury and players, if they have non 0 priors concerning* $\mathfrak{h}_0$ *will converge on their beliefs about* $\mathfrak{h}_0$.

Assume all players i and j with fixed types have distinct but non 0 prior probabilities, $0 \neq \xi_i(\mathfrak{h}_0) \neq \xi_j(\mathfrak{h}_0) \neq 0$, in an ME truth game $\mathcal{G}$. First we look at the dynamics from one stage of play to the next. As the play ρ in $\mathcal{G}$ develops, after every prefix ρ^j that provides evidence that 0 cannot rebut against $\mathfrak{h}_0^j$ constructed over the prefix ρ^j, $\xi_0^{\rho^j}(\mathfrak{h}_0) < \xi_0^{\rho^{j-1}}(\mathfrak{h}_0)$. More generally, for player i or the Jury (call this k), $\xi_k^{\rho^j}(\mathfrak{h}_0) < \xi_k^{\rho^{j-1}}(\mathfrak{h}_0)$ is adjusted similarly as i and the Jury are cognitively skilled and anti-luck adept and update their beliefs about interpretations similarly. Similar observations hold for a play ρ_j that supports $\mathfrak{h}_0$ or refutes a counterargument; in this case, $\xi_k^{\rho^j}(\mathfrak{h}_0) > \xi_k^{\rho^{j-1}}(\mathfrak{h}_0)$.

Imagine now that 0 has a winning strategy to guarantee the condition in equation (11). So for some m and all $n \geq m$, $\xi_{i,j}^{\rho^n}(\mathfrak{h}_0)$ is almost everywhere monotone increasing as the unanswered counterarguments become fewer and fewer but bounded by 1 and above some degree of acceptance α. Thus, for i, j, $\limsup_{n\to\infty} \xi_{i,j}^{\rho^n}(\mathfrak{h}_0)$ exists. And under updating, $\xi_i^{\rho^n}(\mathfrak{h}_0) \to \xi_j^{\rho^n}(\mathfrak{h}_0)$ as $n \to \infty$. Setting $\xi_i^{\rho}(\mathfrak{h}_0) = \limsup_{n\to\infty} \xi^{\rho^n}(\mathfrak{h}_0)$, the beliefs in $\mathfrak{h}_0$ of every player in $\mathcal{G}$ converge to that limit.

Now suppose that 0 does not have a winning strategy. Then for some m and all $n \geq m$, $\xi_i^{\rho_n}(\mathfrak{h}_0)$ is monotone decreasing, bounded and below α. So $\liminf_{n\to\infty} \xi_i^{\rho_n}(\mathfrak{h}_0) = \xi_i^{\rho}(\mathfrak{h}_0)$ exists and each player's belief in $\mathfrak{h}_0$ will converge to this limit. In sum, given that 1 is cognitively skilled, if there are flaws in $\mathfrak{h}_0$, she will find them and this will lead both the Jury and 0 to modify their beliefs accordingly. The uncovered flaws will lead all to either collectively reject $\mathfrak{h}_0$ or to modify it so that in the end all cognitively skilled players fail to find fault with it. If there are no flaws that 1 can uncover or that 0 cannot answer to satisfactorily, then all cognitively skilled participants in G will converge to accepting $\mathfrak{h}_0$.

There are some consequences also for a $\mathfrak{h}_X^G$ that is part of a winning strategy for 0 in an ME truth game G. The first property is completeness, an accounting of all or sufficiently many of the facts the history is claimed to cover. If they are not covered, a cognitively skilled player 1 will uncover this incompleteness. In addition, histories that are part of a winning strategy of 0 in an ME truth game should have predictive and explanatory adequacy.

Definition 16 (Predictiveness). *A history* $\mathfrak{h}_{0,X}$ *over a set of facts X in an ME game* $\mathcal{G}$ *is predictive just in case when* 0 *is presented with a set of facts Y relevantly similar to X, there is an ME truth game* $\mathcal{G}$*' where* $\mathfrak{h}_{0,X}$ *has a consistent extension* $\mathfrak{h}_{0,Y}$ *in* $\mathcal{G}$*' and* 0 *has a winning strategy in* $\mathcal{G}'$.

Definition 17 (Explanatory adequacy). *A history $\mathfrak{h}_{0,X}$ of a truth game $\mathcal{G}$ over a set of facts X is explanatorily adequate just in case if* 1 *asks in* $\mathcal{G}$ why ϕ?*, where ϕ is some proposition about X, then $\mathfrak{h}_{0,X}$ either furnishes an answer to the question or is compatible with an answer that satisfies a cognitively skilled* 1.

Observation 9 *If* 0 *has a winning strategy in an ME truth game $\mathcal{G}$, her history $\mathfrak{h}_{0,X}$ will be predictive and explanatorily adequate.*

Suppose in an ME truth game G that $\mathfrak{h}_{0,X}$ for a set of facts X is not predictive or explanatorily adequate. The cognitively skilled opponent 1 will bring these flaws to light for the Jury. Since the Jury is also cognitively skilled, these flaws will weigh negatively in their evaluation of $\mathfrak{h}_{0,X}$ and lead to a negative evaluation of $\mathfrak{h}_{0,X}$. Hence, no winning strategy for 0 in G.

As an ME truth game is win-lose, if the winning condition is Borel definable, it will be determined [9]; either 0 has a winning strategy or 1 does. Whether 0 has a winning strategy in an ME truth game or not is important: if she does, there is a method for finding an optimal history in the winning set; if she doesn't, an optimal history from the point of view of a truth-seeking goal in the ME truth game is not always attainable. What is needed for 0 to have a winning strategy in an ME truth game? When the facts X that her history is supposed to relate are sufficiently simple or sufficiently unambiguous in the sense that they determine just one history and 0 is effectively able to build and defend such a history. Very simple cases like establishing in the morning whether your daughter has a snack for after school or not are easy to determine, and the history is equally simple, once you have the right evidence: yes she has a snack, or no she doesn't. You can even put a bound on the length of play needed to establish a win. A text which is unambiguous similarly determines only one history, and linguistic competence should suffice to determine what that history is. In general whether or not a player has a winning strategy will depend on the structure of the optimal history targeted, as well as on the resources and constraints on the players in an ME truth game. In future work, we plan to analyze in detail complicated real life linguistic examples like the ongoing debate about climate change where there is large scale scientific agreement but where disagreement exists because of distinct winning conditions.

In the general case, whether 0 has a winning strategy in an ME truth game is highly non trivial. An ME truth game suggests a Peircean "best attainable" conception of truth: it is an "internal" notion of truth based on good epistemic practices and on explanatory and predictive power. But this internal notion also aims at an external ideal. 0's goal in an ME truth game $\mathcal{G}_X$ is finally to reflect the reality about some set of facts X as accurately as possible. To formalize this more external view of truth, assume the Jury has in its possession in an *ME external truth game* $\mathcal{G}_X$ a *ground truth structure* or class of such structures $\mathcal{S}_X$. $\mathcal{G}_X$ is just like an ME truth game except the winning condition in $\mathcal{G}_X$ is that $\mathfrak{h}_0$ converges to determining a structure that is isomorphic to some element of $\mathcal{S}_X$. Building such isomorphic structures can be extremely complex—even Borel complete [54]. Even a winning condition for an ME persuasion game W in which 0 wins by simply replying to every attack by 1 on her history while remaining consistent in an infinite ME game is at least as complex as the procedure for constructing an isomorphic copy of a dense linear ordering. While any winning strategy for the condition in (11)

will suffice to guarantee a win in W, it will not necessarily suffice to guarantee a win in an external truth game.

Observation 10 *Given an ME truth game G_X and ME external truth game $\mathcal{G}_X$, a winning strategy for G_X does not entail a winning strategy for $\mathcal{G}_X$.*

As proof, pick an X with a ground truth structure $\mathcal{S}$ such that finding an isomorphic copy of $\mathcal{S}$ is strictly harder than the complexity of the winning condition in (11).

ME truth games are infinitary games, in which getting a winning history is something 0 may or may not achieve in the limit. There are also finite variants of an ME truth game, in which the winning condition would be to have a 0 defensible history for a certain finite time. In practice, this is what happens in most fields; after a certain point a history about some set of facts (these could be scientific results!) become accepted as true even if there is always the possibility of discovering some flaw in the account. We can also investigate discounted ME games [2], in which the scores assigned to individual discourse moves integrate some discounting factor, to investigate cases where getting things right enough early on in an ME truth game is crucial.

5 Prior work on bias

We've already discussed at length the importance of prior work on ME games and discourse structure for the theory of interpretive bias. We have in addition mentioned work on bias and lexical choice as well as relevant work on dog whistle content. Bias has also been studied in cognitive psychology and empirical economics [57, 58, 60, 59, 25, 33, 34, 32, 64]. Since the seminal work of Kahneman and Tversky and the economist Allais, psychologists and empirical economists have provided valuable insights into cognitive biases in simple decision problems and simple mathematical tasks [13]. Statisticians have also done important work on sample bias.[23]

The bias of framing effects [58], is directly relevant to our theory of interpretive bias. A situation is presented using certain lexical choices that lead to different "frames": $x\%$ of the people will survive if you do z (frame 1) versus $y\%$ of the people will die if you do z (frame 2). In fact, $x + y = 100$, the total population in question; so the two consequents of the conditionals are equivalent. But each frame elaborates or "colors" z in a way that affects an interpreter's evaluation of z. To connect this with ME games, consider two ME evaluation games $\mathcal{G}_1$ and $\mathcal{G}_2$, where 0 makes the play in frame 1 and the play in frame 2 respectively. Imagine now that 1 must assign a type to 0; lexical choice in $\mathcal{G}_1$ will lead 1 to assign 0 a type that seeks to show a benefit of doing z, t_0^b, while in $\mathcal{G}_2$, lexical choice will lead 1 to assign 0 a type t_0^{rsk} of seeking to point out a risk. Following the sort of reasoning in Section 4.1, we can easily see how 1 might give a "Like" to one history and a "Nay" to the other history.

Psychologists, empirical economists and statisticians have typically investigated cases of cognitive bias in which subjects *deviate* from prescriptively rational or independently given objective outcomes in quantitative decision making and frequency estimation, even though they arguably have the goal of seeking an optimal or "true"

[23] See https://www.elen.ucl.ac.be/esann/index.php?pg=specsess#biasesbigdata.

solution. Our analysis of interpretive bias leaves open the question whether there is an objective norm or not. For persuasion games an agent may have a defensible history in the eyes of the Jury even if she is not interested in that norm. Truth games, on the other hand, investigate that norm, whether it is attainable, and, if so, under what conditions.[24]

Our approach to interpretive bias generalizes to a more general view of bias outside of linguistically expressed contents. Understood abstractly, bias is a parameter $\mathfrak{B}$ in a function (a learner if you will) f from X to models or representations of data Y. The models Y may range from simple binary classification schemes (distinguishing, for instance, spam emails from non spam) to representational structures like the histories we have considered as the outputs of ME games. $\mathfrak{B}$ itself typically depends on a number of parameters θ. Thus, our learner abstractly, looks like this: $f_{\mathfrak{B}(\theta)} : X \to Y$. As for the parameters θ, Bayesian inference, for example, which underlies many powerful models of inference and machine learning, [38], defines biases with parameters like: the prior probability distribution over states, which parameters are probabilistically independent, and what kind of conditional probability distribution each parameter abides by (e.g., normal distribution, noisy-or, bimodal). Our analysis of interpretive bias fits into this picture: our input X consists of plays while the output Y consists of full histories or FLFs; our parameters θ are the types for the Jury and the players. [38]'s insights can be applied to the parameters in our analysis of bias.

As in other accounts of bias, ME games allow us to compare biases. We might, for instance, compare biases and the models they generate with "ground truth" as in machine learning or statistics.[25] Further, when ground truth is not independently available, ME games can tell us how to exploit biases optimally.

Our analysis suggests breaking down $\mathfrak{B}_\theta$ into two components: an "authorial" component $\mathfrak{B}_0(\theta_0)$, with θ_0 the relevant parameters for 0's construction of a play, and an "interpreter" component $\mathfrak{B}_1(\theta_1)$ that fills in this play to construct a complete history. Perhaps this division is useful in other analyses of bias as well. In machine learning, some have argued to isolate a "natural" class of examples in the input space X to privilege in learning [41, 42]; this is similar to our authorial input to interpretive bias.

6 Conclusions

In this paper, we have put forward the foundations of a formal model of interpretive bias using epistemic Message Exchange games as developed in [8, 9] and have used these games to model authors' and interpreters' biases. We introduced a particular type of epistemic ME game, an ME evaluation game, to model the strategic consequences of these biases and have linked with important linguistic work on dog whistles to show various ways in which biases can affect the content of a text. Dog whistles are one sort of effect; and what we call semantic bleaching, which to our knowledge has not been investigated in the linguistic literature, is another. We then pursued the strategic consequences of biases and epistemic content with ME persuasion games, a new type

[24] The mathematical structure of ME games also makes it natural to investigate how ME game analyses of bias interact with information-theoretic analyses proposed by [32].

[25] In those fields 'bias' often refers to the divergence between an estimated hypothesis about a parameter and its objective value.

of epistemic ME game that we have proposed here. Finally, we introduced a particular type of ME persuasion games, ME truth games, in which players have optimal biases from an epistemic point of view. The Jury in such a game is we would intuitively call "unbiased."

There are some open questions about bias that ME games can help us answer. ME truth games, for instance, allow us to analyze extant strategies for eliminating bias. Given two histories for a given set of facts, it is a common opinion that one finds a less biased history by splitting the difference between them.[26] This is a strategy perhaps distantly inspired by the idea that the truth lies in the golden mean between extremes, but is this really true? ME games should allow us to encode this strategy and find out.

ME games also introduce new questions about bias. A typical assumption we make as scientists is that rationality would lead us to always prefer to have a more complete and more accurate history for our world. But bias isn't so simple, as an analysis of ME games show. ME games are played for many purposes; non truth-seeking biases that lead to histories that are not a best approximation to the truth may be the rational or optimal choice, if the winning condition in the game is other than that defined in an ME truth game. This has real political and social relevance; for example, a plausible hypothesis is that those who argue that climate change is a hoax are building an alternative history, not to get at the truth but for other political purposes.

Even being a truth interested player can at least initially fail to generate histories that are in the winning condition of an ME truth game. Suppose 0, motivated by truth interest, has constructed for facts $\mathcal{X}$ a history $\mathfrak{h}$ that meets constraints including coherence, consistency, and completeness, and it provides explanatory and predictive power for at least a large subset $\mathcal{Z}$ of $\mathcal{X}$. 0's conceptualization of $\mathcal{X}$ can still go wrong, and 0 may fail to have a winning strategy in interesting ways. First, $\mathfrak{h}$ can mischaracterize $\mathcal{X}$ with high confidence in virtue of evidence only from $\mathcal{Z}$ [40];[27] especially if $\mathcal{Z}$ is large and hence $\mathfrak{h}$ is just simply very "long", it is intuitively more difficult even for truth seeking players to come to accept that an alternative history is the correct one. Second, $\mathfrak{h}$ may lack or be incompatible with concepts that would be needed to be aware of facts in $\mathcal{X} \setminus \mathcal{Z}$. [6, 62] investigate a special case of this, a case of unawareness. To succeed 0 would have to learn the requisite concepts first.

We think epistemic ME games will provide insights into other issues like learning as well. Following [22, 27] learning can be naturally represented as a 0 sum game. An iterated learning process can be represented in an ME game in which 0 makes predictions in virtue of some history, which one might also call a model or a set of hypotheses and for which the winning condition is defined in terms of some function of the scores at each learning round or in terms of some global convergence property. In an ME truth game, 1 can successfully defend a history $\mathfrak{h}_1$ as long as either (a) 0 cannot convince the Jury that her history $\mathfrak{h}_0$ is the right history or (b) 1 can justify $\mathfrak{h}_1$ as an alternative interpretation. In case (a) we say that 0's history $\mathfrak{h}_0$ is not learnable and in case (b) not uniquely learnable. Thus, ME games open up an unexplored research area of *unlearn-*

[26] For instance see, http://www.edu.gov.mb.ca/k12/cur/socstud/foundation_gr9/blms/9-1-3g.pdf.

[27] This option encapsulates the problem of optimizing the decision to exploit a bias that has a certain "local" optimality or to explore the space of possible biases further. There is a large body of literature on this issue [63, 39, 12, 20, 10, 28].

able histories. Consider the bias of a hardened climate change skeptic as player 1: ME games can model the fact that simply presenting new facts to him will not induce him to change his history, even if to a disinterested Jury his history is clearly not in his winning condition. He may either simply refuse to be convinced because he is not truth interested, or because he thinks his alternative history can explain all of the relevant data just as well as a climate science history. For the climate change skeptic, the scientific history about climate change may be unlearnable.

References

1. Afantenos, S., Kow, E., Asher, N., Perret, J.: Discourse parsing for multi-party chat dialogues. In: Empirical Methods in Natural Language Processing. pp. 928–937. Association for Computational Linguistics (2015)
2. Asher, N., Paul, S.: Evaluating conversational success: Weighted message exchange games. In: Hunter, J., Simons, M., Stone, M. (eds.) 20th workshop on the semantics and pragmatics of dialogue (SEMDIAL). New Jersey, USA (July 2016)
3. Asher, N.: Reference to Abstract Objects in Discourse. Kluwer Academic Publishers (1993)
4. Asher, N., Hunter, J.: Interpretive blindness and the impossibility of learning from testimony. In: Proceedings of AAMAS 2021 (2021)
5. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge University Press (2003)
6. Asher, N., Paul, S.: Conversations and incomplete knowledge. In: Proceedings of Semdial Conference. pp. 173–176. Amsterdam (December 2013)
7. Asher, N., Paul, S.: Language games. In: Amblard, M., Groote, P., Pogodalla, S., Retoré, C. (eds.) Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996–2016). pp. 1–17. LNCS, Springer, Berlin Heidelberg (2016)
8. Asher, N., Paul, S.: Strategic conversation under imperfect information: epistemic Message Exchange games. Logic, Language and Information **27.4**, 343–385 (2018)
9. Asher, N., Paul, S., Venant, A.: Message exchange games in strategic conversations. Journal of Philosophical Logic **46.4**, 355–404 (2017), http://dx.doi.org/10.1007/s10992-016-9402-1
10. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Machine learning **47**(2-3), 235–256 (2002)
11. Aumann, R.J.: Agreeing to disagree. The Annals of Statistics **4**(6), 1236–1239 (1976)
12. Banks, J.S., Sundaram, R.K.: Switching costs and the gittins index. Econometrica: Journal of the Econometric Society **62**, 687–694 (1994)
13. Baron, J.: Thinking and deciding. Cambridge University Press (2000)
14. Battigalli, P.: Rationalizability in infinite, dynamic games with incomplete information. Research in Economics **57**(1), 1–38 (2003)
15. Beaver, D., Stanley, J.: Toward a non-ideal philosophy of language. Graduate Faculty Philosophy Journal **39**(2), 503–547 (2018)
16. Benamara, F., Asher, N., Mathieu, Y.Y., Popescu, V., Chardon, B.: Evaluation in discourse: a corpus-based study. Dialogue and Discourse **7**(1), 1–49 (2016)
17. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational linguistics **22**(1), 39–71 (1996)
18. Besnard, P., Hunter, A.: Elements of argumentation, vol. 47. MIT press Cambridge (2008)
19. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of" bias" in nlp. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. p. 5454–5476. Association for Computational Linguistics (2020)

20. Burnetas, A.N., Katehakis, M.N.: Optimal adaptive policies for markov decision processes. Mathematics of Operations Research **22**(1), 222–255 (1997)
21. Burnett, H.: Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. Journal of Sociolinguistics **21**(2), 238–271 (2017)
22. Cesa-Bianchi, N., Lugosi, G.: Prediction, learning, and games. Cambridge university press (2006)
23. Chaterjee, K.: Concurrent games with tail objectives. Theoretical Computer Science **388**, 181–198 (July 2007)
24. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence **77**(2), 321–357 (1995)
25. Erev, I., Wallsten, T.S., Budescu, D.V.: Simultaneous over-and underconfidence: The role of error in judgment processes. Psychological review **101**(3), 519 (1994)
26. Franke, M.: Signal to act: Game theory in pragmatics. Ph.D. thesis, Universiteit van Amsterdam (2009)
27. Fudenberg, D., Levine, D.K.: The theory of learning in games. MIT press (1998)
28. Garivier, A., Cappé, O.: The kl-ucb algorithm for bounded stochastic bandits and beyond. In: COLT. pp. 359–376 (2011)
29. Glazer, J., Rubinstein, A.: On optimal rules of persuasion. Econometrica **72**(6), 119–123 (2004)
30. Harsanyi, J.C.: Games with incomplete information played by "bayesian" players, parts i-iii. Management science **14**, 159–182 (1967)
31. Henderson, R., McCready, E.: Dogwhistles and the at-issue/non-at-issue distinction. Published on Semantics Archive (2017)
32. Hilbert, M.: Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. Psychological bulletin **138**(2), 211 (2012)
33. Hintzman, D.L.: Minerva 2: A simulation model of human memory. Behavior Research Methods, Instruments, & Computers **16**(2), 96–101 (1984)
34. Hintzman, D.L.: Judgments of frequency and recognition memory in a multiple-trace memory model. Psychological review **95**(4), 528 (1988)
35. Hobbs, J.R., Stickel, M., Appelt, D., Martin, P.: Interpretation as abduction. Artificial Intelligence **63**(1–2), 69–142 (1993)
36. Khoo, J.: Code words in political discourse. Philosophical Topics **45**(2), 33–64 (2017)
37. Konek, J.: Probabilistic knowledge and cognitive ability. Philosophical Review **125**(4), 509–587 (2016)
38. L Griffiths, T., Kemp, C., B Tenenbaum, J.: Bayesian models of cognition. Carnegie Mellon University (2008)
39. Lai, T.L., Robbins, H.: Asymptotically efficient adaptive allocation rules. Advances in applied mathematics **6**(1), 4–22 (1985)
40. Lakkaraju, H., Kamar, E., Caruana, R., Horvitz, E.: Discovering blind spots of predictive models: Representations and policies for guided exploration. arXiv preprint arXiv:1610.09064 (2016)
41. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: Unjustified counterfactual explanations. arXiv preprint arXiv:1907.09294 (2019)
42. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: Unjustified classification regions and counterfactual explanations in machine learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 37–54. Springer (2019)
43. Lauscher, A., Glavaš, G.: Are we consistently biased? multidimensional analysis of biases in distributional word vectors. arXiv preprint arXiv:1904.11783 (2019)

44. Lewis, D.: Convention: A Philosophical Study. Harvard University Press (1969)
45. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: A framework for the analysis of texts. International Pragmatics Association Papers in Pragmatics **1**, 79–105 (1987)
46. May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R.: On measuring social biases in sentence encoders. arXiv preprint arXiv:1903.10561 (2019)
47. Mitchell, T.M.: The need for biases in learning generalizations. Department of Computer Science, Laboratory for Computer Science Research ... (1980)
48. Moss, S.: Epistemology formalized. Philosophical Review **122**(1), 1–43 (2013)
49. Parikh, P.: Communication and strategic inference. Linguistics and Philosophy **14**, 473–514 (1991)
50. Parikh, P.: The Use of Language. Stanford University: CSLI Publications (2001)
51. Parikh, P.: Pragmatics and games of partial information. In: Benz, A., Jäger, G., van Roiij, R. (eds.) Game Theory and Pragmatics, pp. 101–122. Palgrave MacMillan (2006)
52. Potts, C.: The logic of conventional implicatures. Oxford University Press Oxford (2005)
53. Saul, J.: Dogwhistles, political manipulation, and philosophy of language. New work on speech acts **360**, 84 (2018)
54. Sobot, B.: Games on Boolean Algebras. Ph.D. thesis, University of Novi Sad, Serbia (2009)
55. Sperber, D., Wilson, D.: Relevance. Blackwells (1986)
56. Stanley, J.: How propaganda works. Princeton University Press (2015)
57. Tversky, A., Kahneman, D.: Availability: A heuristic for judging frequency and probability. Cognitive psychology **5**(2), 207–232 (1973)
58. Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. Science **211**(4481), 453–458 (1981)
59. Tversky, A., Kahneman, D.: Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological review **90**(4), 293 (1983)
60. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. In: Arkes, H., Hammond, K. (eds.) Judgment and decision making: An interdisciplinary reader, pp. 38–55. Cambridge University Press (1986)
61. van Rooij, R.: Signalling games select horn strategies. Linguistics and Philosophy **27**, 493–527 (2004)
62. Venant, A.: Structures, Semantics and Games in Strategic Conversations. Ph.D. thesis, Université Paul Sabatier, Toulouse (2016)
63. Whittle, P.: Multi-armed bandits and the gittins index. Journal of the Royal Statistical Society. Series B (Methodological) **42.2**, 143–149 (1980)
64. Wilkinson, N., Klaes, M.: An introduction to behavioral economics. Palgrave Macmillan (2012)
65. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. Neural computation **8**(7), 1341–1390 (1996)
66. Wolpert, D.H., Macready, W.G.: No free lunch theorems for search. Tech. rep., Technical Report SFI-TR-95-02-010, Santa Fe Institute (1995)