Received: 30 June 2022 | Revised: 22 March 2023 | Accepted: 2 May 2023

DOI: 10.1111/mila.12460

SUBMITTED ARTICLE

WILEY

The rejection game

Luca Incurvati[1] | **Giorgio Sbardolini**[2]

[1]ILLC and Philosophy Department, University of Amsterdam, Amsterdam, Netherlands

[2]MCMP, LMU Munich, Munich, Germany

Correspondence
Giorgio Sbardolini, MCMP, LMU Munich, Ludwigstr. 31, D-80539 Munich, Germany.
Email: giorgio.sbardolini@lrz.uni-muenchen.de

Funding information
Horizon 2020 Framework Programme, Grant/Award Number: 758540; HORIZON EUROPE Marie Sklodowska-Curie Actions, Grant/Award Number: 101064835

We introduce the rejection game, designed to formalize the interaction between interlocutors in a Stalnakerian conversation: a speaker who asserts something and a listener who may accept or reject. The rejection game is similar to other signalling games known to the literature in economics and biology. We point out similarities and differences, and propose an application in linguistics. We uncover basic conditions under which the Gricean maxim of quality emerges from incentives among the players, providing evidence for a functionalist understanding of the Gricean program.

KEYWORDS
computational pragmatics, conventions, quality, rejection, signalling, truthfulness

1 | STALNAKERIAN CONVERSATIONS

Robert Stalnaker's (1978, 1999, 2002) influential theory of assertion has been applied to many issues in philosophy and linguistics, including in semantics, pragmatics, and epistemology. Stalnaker's theory of assertion is based on the notion of a common ground: the set of presuppositions shared by the participants in a conversation. Conversations are communal inquiries, whose goal is to expand the amount of information stored in the common ground. Interlocutors may achieve such a goal via the speech act of assertion.

> [T]he essential effect of an assertion is to change the presuppositions of the participants in the conversation by adding the content of what is asserted to what is presupposed. This effect is avoided only if the assertion is rejected. (Stalnaker, 1978, p. 86)

An assertion is a proposal to update the common ground. The proposal may be accepted, in which case it becomes shared presupposition, or it may be rejected as false, misleading, offensive, or for other reasons still. Stalnaker's observation has appeared to many to capture something fundamental about the function of assertion in discourse.

Stalnakerian conversations have a distinctive decision-theoretic profile, which in this article we set out to investigate from a game-theoretic perspective. There are at least two interlocutors, a speaker and a listener, each with an action to take. The first contributes information by putting forward an utterance among many possible alternative utterances, and the second accepts or rejects it. This interaction may be modelled by a player sending a signal which influences the choice of a second player, who, in turn, determines outcomes for both.

In Section 2, we introduce the game-theoretic structure of a Stalnakerian conversation, which we call a *rejection game*. As we shall see, games of this kind are not unknown to game theorists, nor to other areas of application of game theory, such as economics and biology. A first contribution of our discussion are the distinctions and refinements we introduce to account for the linguistic application we recommend. In Section 2, we also discuss some interesting equilibria of the rejection game, including an optimal equilibrium in which the speaker always tells the truth and the listener always accepts their assertions: a conversational regime in which Grice's (1975) maxim of quality is observed. Focusing on this quality equilibrium, in Section 3, we present our second contribution, which is a study of how quality may arise out of cognitively undemanding constraints on the players, and for which we provide computational evidence. The conclusion recapitulates our findings and discusses their significance for Gricean pragmatics. In the rest of this first section, we lay out some philosophical preliminaries to our study.

1.1 | Signalling games

On our analysis, Stalnakerian conversations are kinds of signalling games. The use of signalling games for the study of language is familiar from David Lewis's (1969, 1972) seminal work on conventions, and much subsequent development. Moreover, the metaphor of conversations as (baseball) games is the basis for Lewis's influential work on pragmatic interpretation (Lewis, 1979).

In a signalling game, signals encode information that can be used to coordinate. Lewis showed that coordination by signalling can be achieved even by agents who do not attach meaning to a signal prior to the interaction. He then used this observation to formulate a hypothesis about the origin of linguistic conventions as equilibria of signalling games—see Skyrms (2010) for a recent development of this hypothesis. Our work is in this tradition, but our aim is not to explain the emergence of language writ large. Our aim is to explain the emergence of conversational conventions, in particular the maxim of quality.

Stalnakerian conversations do not have the structure of Lewisian signalling games for two related reasons. First, in a Lewisian game, the signal receiver reacts by taking an action among at least two, but in principle infinitely many. In contrast, the choice for the listener in a Stalnakerian conversation is more constrained: accept or reject. This matters because (and this is the second reason) rejection has a specific role in discourse. Intuitively, rejection is there to make sure that the speaker behaves appropriately: in a Stalnakerian conversation, the speaker's overt preference is that the listener accepts the assertion—otherwise she would not have it put forward—but the listener prefers to accept what is appropriate and to reject what is inappropriate. Although there are different grounds for speaking and for reacting to speech, for present

purposes we will simplify the discussion and assume that "appropriate" signalling is truthful, "inappropriate" is untruthful.

To appreciate the role of rejection, consider an example. Someone intends to buy a new car, and a salesperson wants to sell. They coordinate just in case the transaction is made. The salesperson knows whether the car is high or low quality, and the potential customer would not spend much money on a low-quality car. So, the salesperson sends a signal to the customer: "The car is high quality". A straightforward (but not particularly sophisticated) way to analyse this interaction as a Lewisian signalling game is to say that if the potential customer refuses to buy, coordination fails, and both are worse off since no transaction occurred. In contrast, in a rejection game, we might say that the potential customer is better off if they reject a bad deal.

In a rejection game the interests of the agents are potentially in conflict—say, if the salesperson is dishonest. Moreover, the response from the signal receiver are constrained: accept or reject. After all, there would be no point in rejection if the interests of the interlocutors always coincided. As Stalnaker suggests in the remark quoted above, rejection has a policing function on common ground management. The natural question to ask, to which we will turn in Section 2, is whether the possibility of rejection suffices to ensure that the speaker behaves appropriately. Does rejection entail honesty? The answer, perhaps surprisingly, is no. However, there are plausible qualifications one might add that lead to a positive answer.

1.2 | The maxim of quality

According to H. Paul Grice, conversations are governed by a maxim of quality, prescribing a speaker to assert what is true (Grice, 1989). Quality is usually glossed by two more specific norms (Grice, 1975, p. 46):

> Do not say what you believe to be false
> Do not say that for which you lack adequate evidence

Grice took the maxim of quality to have some sort of foundational status: "other maxims come into operation only on the assumption that this maxim of Quality is satisfied" (Grice, 1975, p. 27). The same foundational status was attributed to quality by Larry Horn (1984), in his reduction of all Gricean maxims to two overarching principles, Q and R.

> The Q Principle (listener-based): Say as much as you can
> The R Principle (speaker-based): Say no more than you must

By Horn's functionalist reduction, on the assumption that quality is observed, regularities in linguistic behaviour depend on the dynamic balance between maximizing informativity (Q) and minimizing effort (R). We find this picture compelling.

However, the idea that quality should be part of an explanation of linguistic behaviour has been challenged. Wilson and Sperber (2002) argue that "language use is not governed by any convention or maxim of truthfulness in what is said" (p. 583), and indeed that "[t]here is a range of apparent counterexamples to the claim that speakers try to tell the truth" (p. 586). The counterexamples span from non-literal uses, such as irony and poetry, to dishonest users, such as politicians and salespeople. Rejecting quality, however, Wilson and Sperber give up on the

whole edifice of Gricean pragmatics and Horn-style functionalist explanations, which we regard as a significant cost.

It is hard to disagree with Wilson and Sperber's observation: untruthful speech is not infrequent. Nonetheless, there is often a presumption of truthfulness on our interlocutors: we are often expected to speak the truth and expect others do so as well. From the perspective of this article, the general question raised by the debate on quality is whether the structure of combined speech acts (assertion, acceptance, and rejection) may be the source of a pragmatic norm governing honest linguistic behaviour. As we shall see, there are simple conditions on which quality may emerge in ordinary Stalnakerian conversations.

We approach the question of quality from a naturalistic perspective: for the sake of generality, we aim to understand pragmatic conventions without strong idealizations about the individuals' rationality. We do not deny that a rationalistic understanding is possible or desirable. However, our conclusions apply to creatures that are, relatively speaking, cognitively unsophisticated. *A fortiori*, our conclusions apply to rational agents as well. As we show in Section 3, quality is a stable equilibrium of the rejection game under a cognitively modest trial and error learning strategy and given some plausible assumptions about speaker and listener. In particular, we will assume (i) that listeners prefer to reject untruthful speech, (ii) that speakers prefer to avoid rejection, (iii) that interlocutors' interests at least partially overlap, and (iv) that the interlocutors learn from their experience. Our final argument is for two more principles to be added to Horn's functionalist understanding of Gricean pragmatics, which correspond to assumptions (i) and (ii).

> The I Principle (listener-based): Reject inappropriate speech
> The A Principle (speaker-based): Avoid rejection

From the I and A principles, under the remaining constraints we mentioned (iii) and (iv), quality may be expected to arise as a convention of Stalnakerian conversations.

2 | A LANGUAGE GAME FOR REJECTION

In a rejection game, a speaker plays first by asserting something. A listener reacts with assent or dissent.

> Speaker: Franz is in Amsterdam.
> Listener: Yes! He is. / No! He is not.

We introduce two constraints on the speaker. First, we assume that there are several possible assertions to choose from regarding the same topic, so that the choice of the speaker is not trivial. Topics are sets of mutually exclusive information states, such as {Franz is in Amsterdam, Franz is not in Amsterdam}, which can be regarded as the question "Is Franz in Amsterdam?" identified by its answers (Groenendijk & Stokhof, 1984; Hamblin, 1973; Roberts, 2012). Such initial question need not be a polar question as in the example, that is one that admits a positive and a negative answer only. If so, the speaker has a potentially large number of proposals available.

Secondly, we assume that the speaker has private information: the speaker observes an information state among the set of information states that are possible answers to the initial

question. Moreover, we assume for simplicity that the speaker is not confused about it. In other words, we set aside the possibility of errors without malice: the case of a speaker who fails to tell the truth despite their best intentions. In the example, the speaker may observe without uncertainty that Franz is in Amsterdam or that he is not. It follows from the two constraints we have assumed that the speaker has at least a choice between providing a truthful assertion and an untruthful one. In more complicated scenarios, the speaker may have a continuum of options: she can be less and less informative, and more and more misleading.

Once a proposal is offered, it is the listener's turn to decide what to do with it. Just as the speaker could be perfectly truthful, perfectly untruthful, or somewhat misleading in different degrees, so the listener could wholeheartedly accept a speaker's proposal, or firmly reject it, or take fine-grained attitudes in between the two extremes. Again for the sake of clarity and simplicity, we set aside a more nuanced pragmatics of acceptance and rejection, and assume a discrete choice. The resulting game-theoretic structure is represented in Figure 1.

More formally, a rejection game consists of a set of players (at least, speaker S and listener L), a set of information states (at least p and q, which we assume to be mutually exclusive), and actions and utility functions for each player. See Figure 2, for a representation of the game in extensive form, in which actions are labelled as "p" and "q" for S and A and R for L. See Figure 3, for a representation of the agents' decisions if Nature selects state p: it is the top side of the larger diagram in Figure 2. In technical terminology, Figure 3 represents the speaker's *subgame* of the rejection game under initial state p. Importantly, here it is not assumed that the listener knows that they are playing under p—as indicated explicitly by the dashed lines in Figure 2 connecting the listener's nodes. Payoffs will be discussed later; for now we assume that $a, b > 0$ and $\lambda \geq 0$.

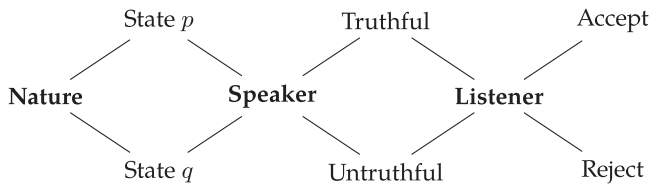

FIGURE 1 Decision tree of a rejection game with two states and two signals.

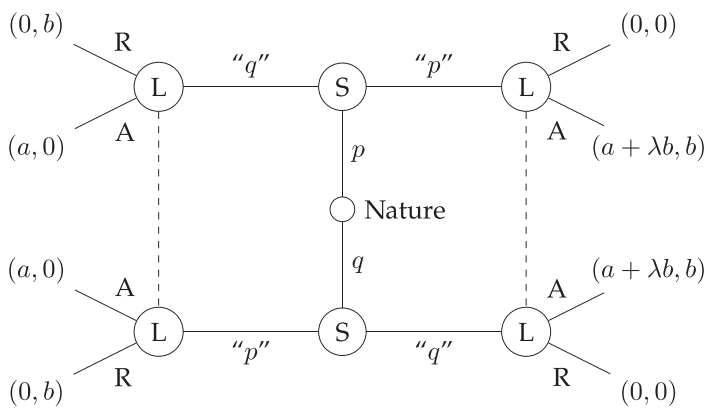

FIGURE 2 Extensive form representation of the rejection game with two states.

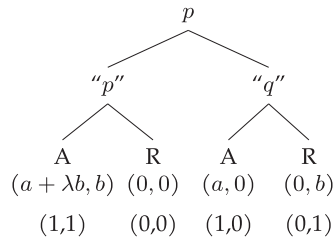

FIGURE 3 Speaker's p-subgame. Numerical values are obtained by $a = b = 1$ and $\lambda = 0$.

It is the job of a "dummy player", Nature, to initialize the game by selecting an information state to be observed by S: this is the true state in which the game takes place. S's strategies are functions from initial states to update proposals, and some possible strategies for S stand out. In particular, S may send "p" if p and "q" if q. We label both actions T for Truthful. Alternatively, S may send "p" if q and "q" if p. We label both actions U for Untruthful.

L has two actions: acceptance A and rejection R. L does not observe the initial state selected by Nature (hence the dotted lines in Figure 2). We may define L's strategies as functions from update proposals to final states. The set of final states is $\{p, q, \bot\}$, with $\bot$ a "dummy state" analogous to a punctuation mark (Tennant, 1999). A is the map from "p" to p and from "q" to q, whereas R is the constant function from update proposals to final states that outputs $\bot$ for any input.

We may think of the players' actions in rejection games as linguistically realized by means of speech acts performed with a certain force: the speaker's assertions are utterances of declarative statements, and the listener's reactions are expressions of force indicators such as *Yes* and *No*. Which final state is selected can then be calculated combining the actions of speaker and listener: $\mathrm{AT}p = p$, that is, p is the update the players are committed to in case they accept a truthful proposal in state p, whereas $\mathrm{RT}p = \bot$, that is, the players are committed to no update in case of rejection. Similarly for other strategy combinations.

The payoffs formalize the intuitive idea that S prefers that L accepts, while L prefers to accept if S is truthful, and to reject otherwise. (The role of λ, a parameter we introduce to refine speaker's behaviour, will be discussed later.) These payoffs rationalize the interlocutors behaviour: if S puts forward an update proposal, her overt goal is for the proposal to become a shared presupposition. In actual cases, speakers may be duplicitous, or engaged in counterfactual reasoning: they may pretend to want a claim accepted, when in fact they do not believe it. Similarly, listeners can pretend to accept, for example for the sake of a *reductio* argument. We follow Stalnaker's (1999, 2002) proviso: these are actions interlocutors perform for conversational purposes, whatever their actual purposes.

As the diagrams show, the choice of the speaker determines which (subsub)game the agents are ultimately playing. If S is truthful, it is in the interest of L to accept. (Of course, L does not know whether S is truthful or not.) Thus, under a T node, both prefer the same action, namely A. Not so if S is untruthful: it is then in the interest of L to reject. Thus under a U node, their payoffs are mutually exclusive. For comparison, consider two types of games that may already be familiar, shown in Table 1 with simple numerical payoffs for illustration.

In a coordination game, the players always prefer the same outcomes. If Column chooses C_1, both Column and Row prefer Row to choose R_1, and vice versa: the interests of Column and Row are perfectly aligned. This is the kind of games used by David Lewis (1969) to study

TABLE 1 Three games.

	C_1	C_2		C_1	C_2		T	U
R_1	1,1	0,0	R_1	1,0	0,1	A	1,1	0,1
R_2	0,0	1,1	R_2	0,1	1,0	R	0,0	1,0
Coordination game			Zero-sum game			Rejection game (speaker's subgame)		

linguistic conventions. In contrast, in a zero-sum game the players always prefer different outcomes: if Column chooses C_1, Column prefers Row to choose R_1 but Row prefers Row to choose R_2. Likewise for the remaining combinations of choices. The players' interests are always in conflict. Zero-sum games always have a winner and a loser (as any sport matches in which there is no tie, like tennis). Compare coordination game and zero-sum game with a speaker's subgame of a rejection game (whether the initial state is p or q). As the table shows, the payoff distribution is neither coordination nor zero-sum. We will discuss some of the equilibria compatible with these payoffs in Section 2.1.

Since the basis for the rejection game is not coordination, the rejection game is not a signalling game in the sense of Lewis (1969). However, there are Lewisian signalling games which approximate the rejection game in various respects. With no ambition to be exhaustive here, there are at least two related lines of research worth mentioning. First, there is the study of linguistic pragmatics. Crawford and Sobel (1982) and Martinez and Godfrey-Smith (2016) study signalling games in which the interests of the players need not align, and the signal is on a continuum from fully informative to completely uninformative. In these generalizations of Lewis's game the possibility of conflict is accounted for by varying degrees of truthfulness. Thus if the speaker has no interest in letting the listener know where Franz is, they can be less than fully informative: "Franz is somewhere in Europe". Signalling games with conflict of interests have been studied in pragmatics and especially game-theoretic pragmatics (Ahern & Clark, 2017; Asher & Lascarides, 2013; De Jaegher & van Rooij, 2014; van Rooij & Sevenster, 2006). In this line of research, the listener's action models the process of disambiguation and interpretation rather than the structure of speech acts in conversation.

Second, there is the study of bargaining in economics—recall the car sale example from the introduction. Several so-called persuasion games have been developed to study the influence of signalling on consumer's choice (Kamenica, 2019; Kamenica & Gentzkow, 2011; Pitchik & Schotter, 1987), and some of these games bear a close similarity to the rejection game—see also Sbardolini (2022) for an application of the theory of persuasion to linguistics. This line of work includes mathematical analyses of equilibrium conditions for games and subgames, on a variety of modelling scenarios. Several of these techniques, extensions, and refinements, could be used to analyse the rejection game and its variants. However, the rational choice approach of formal economics usually requires strong idealizations concerning the intellectual abilities of the players. We will make some such assumptions in Section 2.1, to discuss some notable equilibria of the rejection game, and we regard this as an important avenue for further research. However, for dialectical purposes regarding the debate on quality, we prefer to discuss conversational conventions for relatively unsophisticated agents.

In this section, we have introduced the general structure of a rejection game. We made several assumptions for the sake of simplicity that could be revised in future research: in particular, we have assumed that the speaker forms accurate beliefs about their evidence, that they can only be truthful or untruthful, and that the listener can only accept or reject. It would be

natural to extend the present work by introducing degrees at all levels: the speaker may be more or less confident in what the evidence is, they may approximate the truth, and the listener may want to consider partial acceptance and partial rejection ("If you say so" or "Not exactly"; see the discussion in Horn, 1989, Walker, 1996, and Incurvati & Schlöder, 2017). Degrees in each of these dimensions would account for more fine-grained distinctions, and potentially more phenomena.

Beyond degree-theoretic generalizations of the current framework there are probabilistic extensions, such as Bayesian pragmatics (see Franke & Jäger, 2016, and references therein), which includes models of communication under uncertainty or conflict of interest such as the Rational Speech Act model (Frank, 2017; Franke, 2011; Goodman & Frank, 2016; Sumers et al., 2021). This literature provides a further bridge between the two areas of study we mentioned, interpretation, and persuasion, by offering some insight into the cognitive science of social understanding (Barnett et al., 2022; Goodman & Stuhlmüller, 2013; Oey et al., 2023; Vignero, 2022). We regard it as a strength of a game-theoretic analysis of Stalnakerian conversations that speech act theory can then be seen as central to multiple areas of inquiry. More generally, a probabilistic generalization of the current framework comes with pros and cons. A Bayesian model is expressively more powerful, hence theoretically useful, and more easily related to much ongoing research in cognitive science. However, the fine-grained distinctions thereby achieved must be handled with care: at the very least, it should be confirmed empirically whether probabilistic distinctions are required to explain the observable phenomena, and it should be clarified what constrains priors and likelihoods on pain of unfalsifiability (see Hahn, 2014, for a review of recent debates on Bayesianism). More fine-grained distinctions than we assume lead to natural generalizations of our discussion, and we advocate the virtues of caution in their pursuit.[1]

2.1 | Quality, noise, and the community of liars

The concept of Nash equilibrium offers the simplest answer to the question of what to do in a game. A *Nash equilibrium* is a combination of actions such that no player would be better off by a different choice, assuming everyone else's choices remain the same (Nash, 1950; Osborne & Rubinstein, 1994). We can look at Table 1 for some examples. In the coordination game, there are two *pure* Nash equilibria: C_1R_1 and C_2R_2. In a zero-sum game, there are only *mixed* (or impure) Nash equilibria: for example, Column plays C_1 and Row plays R_1 both with 50% probability.[2]

What about the rejection game? Is it not obvious that the speaker is always truthful and the listener, consequently, always accepts? After all, the speaker wants her assertion to be accepted,

[1]For illustration, we noted above that there are "weak" forms of rejection besides the binary accept/reject distinction we discuss here (Incurvati & Schlöder, 2017). A three-way distinction between accept, reject, and neither, however, is far short of having as many forms of rejection as there are real numbers between 1 (fully accept) to 0 (fully reject). We think that it should be assessed in each specific instance whether the added complexity is, on balance, an asset or a liability.

[2]There are several interpretations of mixed strategies: as the propensity of an agent to perform an action, the frequency of performing it over repeated games, or the uncertainty of the agent as to which action to perform. We need not take a stand on this interpretive issue. In the second part of the article, as we describe a computational model of the rejection game, the frequentist interpretation is perhaps more prominent. For this reason, we use frequentist language throughout.

and the listener accepts just in case the assertion is true. So the speaker can reason that she should send only true messages, that are guaranteed of acceptance.

This is indeed an equilibrium of the rejection game. We may call it the *quality equilibrium*, since the speaker is perfectly truthful and the listener always accepts. Assume the numerical payoffs of Table 1, for both games under p and under q, and keep $\lambda = 0$. Then TA is a Nash equilibrium: neither player can improve on their outcomes if the other's choice remains the same. If the speaker is truthful, the listener has nothing to gain from rejecting every now and then, but only something to lose. Conversely, if the listener accepts, it is pointless for the speaker to lie occasionally: by doing so, the speaker would not be worse-off but also not better-off. (In technical terms, TA is a *non-strict* Nash equilibrium because neither action is strictly better for the agent performing it under all circumstances.).

The argument in the previous paragraph relies on a few implicit assumptions. First, we have assumed, and will assume throughout this section (but drop this assumption later in the article), that the interlocutors know the game they are playing, including which choices they have and their outcomes for both. Idealizations of this kind are commonplace in rational choice theory, and connect formal game-theoretic structures to agents' decisions. Second, we have assumed that the speaker has no particular preference as to which update proposal the listener accepts. This is not always the case, as we have already alluded to in the interaction sketched in the introduction. A salesman wants a customer to buy a car. The initial states are high-quality car and low-quality car. The customer can buy or not, but would rather get the low-quality car for cheap or the high-quality car for an expensive price. The salesman could be truthful, of course. But he might try to get the customer to pay a lot for the low-quality car as well as the high-quality car. In this case, something is at stake for the salesman (S) with regards to the initial states. He prefers that the customer (L) accepts one update proposal over the other, and pays accordingly. Will the salesman tell the truth?

Let us assume that S prefers that L accepts state p over state q. Then, S will send "p" if p is the case, but has no incentive to tell the truth if q is the case. Indeed, S might assert "p" regardless. If so, her signal is not credible (Farrell & Rabin, 1996; Sbardolini, 2022; Sobel, 1985; Stalnaker, 2006): one cannot promise "Lowest Prices Today!" every single day of the year and hope to be believed. Suppose furthermore that S's unilateral preference is known by L.[3] Then, L would do well to accept "p" at a fixed *ratio* r: her prior subjective probability that p. In this scenario, L's choice is no longer influenced by S's assertion, since the decision whether to accept or reject is based on the prior. We may call this scenario the *Noise* equilibrium, since in this outcome S's signal is just noise that does not affect L's decision. Such an equilibrium is not a very good outcome for S, because she only receives a positive payoff with probability r, nor for L, because L cannot exploit S's information to make an informed decision. It is an equilibrium nonetheless: neither can improve their outcomes given what the other does.

There are other possible equilibria, with additional layers of sophistication. Let us keep the assumption that S has a unilateral preference for p over q and that L knows it. Then S is always truthful if p, and can occasionally lie if q, but not as recklessly as in the Noise equilibrium. Instead, let us suppose that S tries to maintain some credibility. For example, assume that S receives utility 1 if "q" is accepted, and $1 + x$ if "p" is accepted. If q is the case, S may decide to be truthful with a certain probability. Then there is what we might call the *semi-truthful* equilibrium: L

[3]What if the speaker's biases are not public information? In this case, the listener might not have knowledge of the speaker's preferences but might still have credences. A natural application of the rejection game in a probabilistic setting is to study interlocutors' behaviour with incomplete information.

always accepts the assertion "q", since S would not assert "q" unless q was the case, and accepts "p" with a probability that is inversely proportional to S's added benefit from acceptance of "p": $1/(1+x)$. Therefore, the more S has to gain from lying, the less likely is the listener to accept. All points at which the speaker cannot gain more from the acceptance of "p" without making it less probable that the listener will accept, and vice versa, are semi-truthful equilibrium points.

There are still more equilibria, which can be found by refining or modifying the assumptions we have made about the interaction and what the interlocutors know about it. The general point is that it is not inevitable that speakers are truthful, just in virtue of participating in a Stalnakerian conversation. In this sense, we are sympathetic to Wilson and Sperber's (2002) emphasis on untruthful speech. Ordinary speakers often tell the truth, but sometimes do not, especially if it is to their advantage: some equilibria of the rejection game show the conditions on which strategic deception is to be expected.

There remains a question about the status of quality as a conversational convention. In many contexts, we do seem to have a tendency towards truth-telling, even though there might be occasional free-riders who exploit a default expectation by getting away with a lie. By contrast, consider an opposite convention: a community in which the norm is to lie, and only on occasion speakers are truthful if it is to their strategic benefit. In the community of liars, if you ask me whether Franz is in Amsterdam (and nothing is at stake for me), I would tell you that he is in Amsterdam when my evidence is that he is not, and that he is not in Amsterdam when my evidence is that he is. In the community of liars, you would do best with constant rejection. This is indeed an equilibrium of the rejection game, whose outcome is that the interlocutors never perform any conversational update: if S is always untruthful, L's best reply is to reject, and conversely if L always rejects, S has no (strictly) better reply than untruthfulness.

The community of liars is very different from our community. Among the liars, if you discovered that I told you the truth about Franz's whereabouts, you would wonder what special reasons I might have had to do so. We do not need special reasons to tell the truth, when nothing is at stake (Abeler et al., 2019). When we discover that someone lied to us, we immediately ask why, not the other way around. The status of quality as a default of ordinary conversations calls for an explanation. The question is then about the conditions for the quality convention to emerge, as opposed to other equilibria of the rejection game.

The community of liars scenario is a notable illustration of the difference between Lewisian signalling games and the rejection game. As we said, such a community is an equilibrium of the latter. In contrast, it is not straightforward how to even describe it from Lewis's perspective. Consider a Lewisian signalling game in which if S observes that p, they assert "q", and vice versa, and if L is told "q", they form the belief that p, and vice versa. It would be mistaken to describe this convention as one in which the speaker always lies and the listener always forms the belief contrary to what they had been told. From the perspective of Lewis (1972), this convention is a language in which *water* means fire and *fire* means water, but the signalling is still reliably correlated with the external environment. Lying is only possible, from Lewis's perspective, on the assumption of such a correlation existing. This difference between Lewisian signalling and the rejection game reflects the focus of the latter on conversations, not on the origins of language itself.

2.2 | An evolutionary argument for quality

There are two families of arguments for equilibrium selection: rational choice arguments and evolutionary arguments. Rational choice arguments tend to involve some pretty sophisticated

meta-reasoning about the game on part of the agents, who are assumed to have knowledge or credences about each others' preferences and their environment, and who reflect on what everyone knows, on what everyone knows that everyone knows, and so on for every iteration of the knowledge operator (*mutatis mutandis* for belief or degree of belief). We have made assumptions of this kind to describe the equilibria of the previous section. However, rational choice arguments are potentially a liability in an explanation of quality, since they ascribe idealized reasoning powers to ordinary interlocutors. Sperber and Wilson (1995, 2002) reject both quality and such idealizations. The list of critics of ideal rationality is indeed much longer and includes Burge (1975), Davidson (1984, 1986), and recently Lederman (2018).

Evolutionary arguments may be found in biological explanations of truthfulness in the animal kingdom. Animal behaviour sometimes reveals information about the agents involved, for example in the case of birds, bees, and monkeys (Haldane, 1992; Hollén & Radford, 2009; Lachmann et al., 2001). An inquiry into the origins of quality among sophisticated animals such as ourselves may well get started from a humble place. Evolutionary arguments do not assume common knowledge of preferences or self-reflection. The intuitive idea is that small discrepancies in the successfulness of an interaction may add up to big differences in the long run, over repeated rounds of the game (Skyrms, 2010).

Many biological interactions are partly competitive, as the rejection game. Unfit males in the wild have reproductive interests that are in conflict with those of the females, who prefer fit partners. The females have to sort a good from a bad match, but all males could in principle send the same signal of fitness. Some of them lie, and get away with it (there are free-riders), but in most cases, the females do select for higher fitness—that is the way of evolution by natural selection. Various hypotheses have been formulated to explain the maintenance of honest behaviour as a default of animal communication. A famous hypothesis is Zahavi's handicap principle (Lachmann et al., 2001; Penn & Számadó, 2019; Zahavi, 1975). Accordingly, honest signalling takes effort, and the benefits of deception do not balance its costs. Developing a long and colourful tail takes good genes, and only a male bird with fitness higher than average can afford it. So the female bird picks on the colourful tail as proxy for fitness.

However, a cost-based account of honest signalling in ordinary conversations is somewhat implausible. To follow the analogy, we would have to posit costs inflicted on truth-tellers, on the assumption that liars cannot bear them. Is telling the truth so hard? Rather than different costs for truthful and untruthful speech, it is perhaps more plausible to imagine that there are different *benefits* attached to telling the truth. (Signalling models with differential costs or differential benefits are to some extent formally equivalent: see Zollman et al., 2013, for an analysis.).

We will present an evolutionary argument for quality on the basis of differential benefits. We will consider the most basic case: the speaker has no preference for one state over the other (unlike in the car sales scenario), and the interlocutors have no knowledge of each others' preferences. That is, we discuss a Stalnakerian conversation with nothing at stake as to whether p or q, between cognitively limited agents who have no knowledge of each others' minds. There are other scenarios which might invite different ways to account for quality, but we focus on this basic case for simplicity and in the hope that, in future research, we might generalize from the simple case.

Even so, there are different ways to explain the benefits of truth-telling. We will assume that ordinary speakers have a preference for honest behaviour, even in situations of potential conflict, because they look at their partners with at least some benevolence or *sympathy* (following David Sally, 2000). Some evidence for sympathy comes from the empirical literature on the ultimatum game (Güth et al., 1982), which bears formal similarities to the rejection game. In the

ultimatum game, a *proposer* has to decide how much of a sum of money x to offer to a *responder*, and offers y with $0 \leq y < x$. The responder can accept the proposal or reject it. If she accepts, the responder gets y, and the proposer keeps the difference $x - y$. If the responder rejects, both get nothing. On the one hand, the Proposer should offer as little as possible: for the responder, even 1 cent is better than nothing, so she should accept 1c. On the other hand, experimental results across different demographic groups by age, culture, and country, consistently show that the vast majority of Proposers come up with roughly fair divisions, for example a 50:50 or 60:40 split (Güth et al., 1982; Roth et al., 1991). Human beings appear to be somewhat sympathetic for their partners (Bicchieri, 2006; Fehr & Schmidt, 1999; Rabin, 1993).

In low stakes conversations, as we are assuming, sympathy plausibly applies. Sally (2000, 2003) emphasises the importance of taking other players' perspectives when playing a language game:

> Sympathy is negative or nil for enemies or distant strangers, slightly positive for acquaintainces and those nearby, and near to one for family and close friends. As a result, the same game may be played quite differently when conducted between friends, when the players can see each other or talk to each other, when they have a shared background, or when their commonality is otherwise emphasized. (Sally, 2003, p.1227).

One thing is to have an assertion accepted, another is to have it accepted in such a way that it goes to the listener's advantage as well. We assume that sympathy introduces benefits for the speaker conditional upon acceptance: so long as the speaker wins, the listener might as well win. In other words, sympathy is a conditional incentive for speaker's cooperation. It is not equivalent to assuming outright cooperativity: there are still outcomes in which the interlocutors have conflicting preferences—otherwise, as we said, there would be no point in rejection. However, a sympathetic speaker has two kinds of preferences: an unconditional preference for the update proposal to be accepted, and a secondary preference for the listener to receive positive payoffs once the proposal is accepted.[4]

We do not claim that human beings are either sympathetic or not: more realistically, in linguistic interactions we sometimes take the perspective of our interlocutors, and sometimes we do not. Lack of sympathy means that the speaker merely prefers that the listener accepts the assertion that was made, whether or not it was true. Thus, a non-sympathetic speaker simply asserts what is likely to be accepted regardless of truth: in Harry Frankfurt's (1986) technical sense, she is inclined to *bullshit*, at least so long as she can get away with it. Therefore, non-zero sympathy is a weak assumption: we do not require that speakers care about the overall well-being of listeners, only that speakers are not bullshitters.

We treat sympathy as a parameter λ that re-distributes a proportion of the listener's payoffs to the speaker. It can be a very small proportion if there is little sympathy among the interlocutors, or it can be significant. In principle, this generates an asymmetry between the speaker's preferred outcomes, which might just be enough for the speaker and the listener to coordinate

[4]Assuming non-zero sympathy generates a trade-off in a scenario such as the car sales example, in which we assume that the speaker has a unilateral preference for one update over the others. Would a salesperson trick even family and close friends into buying low-quality cars for a lot of money? In future work, it would be worth exploring the effects of unilateral speaker preferences in conversations with non-zero sympathy. In such an inquiry, it would be necessary to account for the expectation of sanctions, such as negative repercussions for deceivers, which might affect the decision of a speaker. We set these complications aside.

on the quality equilibrium in the long run.[5] As we shall see, sympathy can indeed be a driving force towards optimization in the selection of communication strategies in rejection games, but a regime of quality—perhaps surprisingly—is not supported solely by the force of the interlocutors' mutual sympathy.

3 | THE MODEL AND THE RESULTS

In this section, we give a formal presentation of the model we used, and we discuss our findings.

3.1 | Implementation

To study Stalnakerian conversations as a rejection game, we let the players play the game multiple times, and allow for their cumulative experience to determine the probability of choosing a course of action. As we discussed, we assume that the initial states are equiprobable and that S has no preference for a state over another (no weights attached to their choice). To implement the effects experience, we assume a simple reinforcement learning algorithm (Herrnstein, 1970; Roth & Erev, 1995), which formalizes the dynamics of learning by trial and error. Cumulative experience affects the probability of making a choice in response to the feedback one gets from making it. Thus, the more the speaker asserts truthfully, the more the listener is likely to trust her, the more the speaker is likely to assert those truths again.

Standard implementations of reinforcement learning induce some biases, which may be corrected by constraints on memory. In other words, there should be a way to discard old information, which otherwise influences the players' choices for disproportionately long time (Spike et al., 2017). Here we employ a version of Moran learning (Moran, 1962), in which we interpret information loss as resulting from the fixed and finite size of the agents' memory, and which is known to facilitate efficient solutions in a number of related games (Barrett, 2009; Huttegger & Zollman, 2011).

In a rejection game with two information states, speaker and listeners have four strategies each. In the following table, TU is the rule to tell the truth if p and lie if q, and the remaining strategies are named accordingly.

(TU)	Always assert "p"	(AR)	Accept "p", reject "q"
(UT)	Always assert "q"	(RA)	Accept "q", reject "p"
(TT)	Assert "p" if p, assert "q" if q	(AA)	Accept everything
(UU)	Assert "p" if q, assert "q" if p	(RR)	Reject everything

The quality equilibrium obtains if TT and AA are selected. The community of liars equilibrium obtains if UU and RR are selected. The question is whether, in the long run, speaker and listener optimize their behaviour toward the quality equilibrium by simple trial and error.

[5]Formally speaking, so long as $\lambda \neq 0$, the quality equilibrium is an *evolutionary stable strategy* of the rejection game (Maynard Smith, 1982). Let eu_i be player i's payoff and, with reference to Figure 2 and the definitions above for T and U, let $a = b \neq 0$ and assume even probabilities among states. Then $eu_S(\mathrm{T},\mathrm{R}) = eu_S(\mathrm{U},\mathrm{R})$ but $eu_S(\mathrm{T},\mathrm{A}) > eu_S(\mathrm{U},\mathrm{A})$ so long as $\lambda \neq 0$. For the listener, $eu_L(\mathrm{A},\mathrm{T}) > eu_L(\mathrm{R},\mathrm{T})$. Hence, $\langle \mathrm{T},\mathrm{A} \rangle$ is an evolutionarily stable combination of strategies.

If S is an initial state (either p or q), and s' a final state (either p or q or $\perp$), payoffs for speaker and listener are defined with $a, b > 0$ and $\lambda \geq 0$.

$$u_S(s,s') = \begin{cases} a+\lambda b & \text{if } s = s' \\ a & \text{if } s \neq s' \text{ and } s' \neq \perp \\ 0 & \text{if } s \neq s' \text{ and } s' = \perp \end{cases} \quad u_L(s,s') = \begin{cases} b & \text{if } s = s' \\ 0 & \text{if } s \neq s' \text{ and } s' \neq \perp \\ b & \text{if } s \neq s' \text{ and } s' = \perp \end{cases}$$

The speaker is better off either if the final state matches the initial information ($s = s'$), in which case she gets her acceptance payoff (a) and a boost proportional to her level of sympathy λ, or if $s' \neq \perp$, as her assertion has then been accepted. The listener is better off either if $s = s'$, as she has accepted something true, or if the final state is $\perp$ and the speaker's assertion does not correspond to the initial evidence, as she has then rejected something false.

An interaction is *successful* if and only if the speaker is truthful and the listener accepts for every state chosen by Nature at the beginning of a round. We calculate the probability of success at a time n as the sum over states S and update proposals t of the product of the probabilities that a true assertion is sent and accepted given the state at time n. If the probability of success tends to 1, the speaker tends to play T and the listener A in all states.

$$\sum_s \sum_t P_S(s|s)(n) \times P_L(t|t)(n) \quad \text{Probability of success}$$

A round of the game determines how the agents will play at the next round. The payoff assignments above define a deterministic learning dynamics for speaker and listener. We specify, for each player, what the initial accumulated rewards are, what the probability of making a move is, and how rewards accumulate. The learning hypothesis we are assuming is that the players have a limited memory of their accumulated rewards and choose to act based on their memory.

Let $V_i(n,x,s)$ be player i's reward at time n for playing x in state S. Initially, all speaker's choices and all listener's choices have equal accumulated rewards. Thus before the first round of play, we assume that for all x and S, $V_i(0,x,s) = k$ for some $k > 0$. The probability that an agent chooses x in S at time n is given by the ratio between the rewards for x and the total rewards.

$$P_i(x|s)(n) = \frac{V_i(n,x,s)}{\sum_y V_i(n,y,s)}$$

Rewards accumulate by an Update condition and a Replacement condition given the utilities u_S and u_L defined above (in the computer model, we set $a = b = 1$ and $r = 1$ in the Replacement condition below).

$$V_i(n+1,x,s) = V_i(n,x,s) + u_i(s,s') \quad \text{Update}$$

$$V_i(n+1,x,s) = \begin{cases} V_i(n,x,s) - r & \text{if } V_i(n,x,s) > 1 \\ V(n,x,s) & \text{otherwise} \end{cases} \quad \text{Replacement}$$

An agent is rewarded for playing x in S proportionally to her utilities: this is the Update condition. If the agent's choice of action was successful rewards accumulate; otherwise $u_i(s,s') = 0$, hence they do not. Moreover, if $\lambda \neq 0$ then honest behaviour is learned faster. Finally, at each round of play, an action x is picked at random from those available to the agent, and (provided there are some accumulated rewards for it, to prevent the probability of playing x from going to 0) its rewards are diminished slightly. This is the Replacement condition. It has the consequence that the sum of rewards for all actions remains (roughly) constant. Intuitively, the replacement condition prevents the agent's memory of past interactions from growing indefinitely. (Of course, at a given round, Replacement might undo the reinforcement added by the Update. While there is a small chance of this, it does not matter in the long run.).

3.2 | Results

By simple chance, the speaker picks TT one in four times (25%). The most important finding of our study is that, while sympathy without rejection and rejection without sympathy scarcely affect speaker's behaviour, truthfulness becomes a mathematical certainty with rejection and sympathy together: the players rapidly converge on quality in all games with two and three information states. With four, the speaker's chance to select perfect truthfulness is one in 256 ($= 0.392\%$). However, perfect truthfulness is still selected more than 90% of the times by a sympathetic speaker whose interlocutor can reject. The combination of sympathy with the possibility of rejection makes quality the norm. The Appendix includes a report of our results.

Hence, speaker and listener can find the quality equilibrium by simple reinforcement learning, while participating in an interaction that has the cognitively undemanding properties we have described—without sophisticated self-reflection, nor reasoning about each other's knowledge, nor about each other's preferences. Yet the probability of success goes to 1 if the speaker is somewhat sympathetic, and the listener can reject. Eventually, the speaker's search space for the perfectly truthful strategy becomes too vast (with five or more states, for example): she cannot learn to be truthful fast enough, and the listener gains more for always rejecting the frequent falsehoods. This effect is due to an expected computational limit of our model.

Interestingly, we have found that rejection alone and sympathy alone do not achieve much: see Figure 4. A speaker who may face rejection but has no sympathy has roughly one in four chances of playing TT (given two initial states) after repeated rounds of the game: such a speaker selects TT with approximately 22% probability, not far from simple chance. The function of rejection is to block an update, but its effect is not to impose a regime of truthfulness on the conversation.

Sympathy, without rejection, does not drive Stalnakerian conversations to the quality equilibrium either. Such a speaker selects TT with approximately 27.6% probability, with two states—again not far from the chance baseline. The reason is probably that S's choice is invariably reinforced: she simply picks a strategy at the outset, and essentially gets confirmed in her choice with only some noise generated by the learning process. In this case, the probability of success is simply determined by the probability that a speaker randomly picks the right strategy.

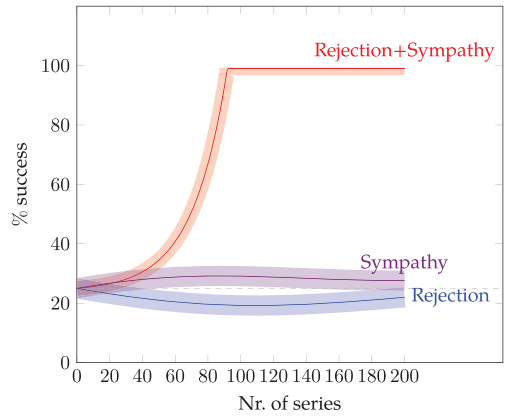

FIGURE 4 Smoothed average likelihood of truthfulness (for two states and $\lambda = 1$).

This is evidence that the sympathy assumption is quite weak and in fact does not amount to a stipulation in favour of quality.

Why does the combination of rejection and sympathy work then? In simulations, sympathy works by guiding the speaker towards truthfulness, even slightly, and rejection works by contrasting the speaker's learning inertia. Without rejection, the speaker quickly settles into familiar patterns of behaviour, even if these might be sub-optimal. With rejection, the speaker takes much longer to settle on a course of action, because her choices that do not favour the listener tend to be filtered out. This gives the speaker time to explore many possible strategies, trying out more options. So it is more likely that the speaker eventually stumbles upon the utility-maximizing strategy TT. Once the speaker has found it, quality quickly develops.

Summarizing, we would like to highlight two aspects of our results. First, we presented a countermodel to the hypothesis that rejection alone is sufficient to establish truthfulness in a Stalnakerian conversation. This is perhaps surprising, since one might have thought this hypothesis to have some initial plausibility. However, it does not hold under the simple scenarios we have considered. Second, we presented a model in favour of the hypothesis that a marginal amount of sympathy among interlocutors is sufficient to establish truthfulness in a Stalnakerian conversation. We have argued that sympathy is cognitively plausible, and we have shown that its effects are limited, since sympathy alone (that is, without rejection) is again insufficient for truthfulness.

4 | CONCLUSION

This article makes two main contributions. First, we introduced rejection games, formalizing the decision-theoretic structure of a Stalnakerian conversation, in which the speaker provides an update proposal that the listener may accept or reject. We have compared these games to Lewis/Skyrms signalling games, already familiar in linguistics, and to other games known in Economics and Biology. We pointed out several directions for further research, by modifying assumptions we made along the way. Second, we formulated a hypothesis about the origins of a convention for truthfulness in conversation, showing that it is the natural result of linguistic interactions that have the structure of rejection games, with sufficient sympathy among the

players, under modest cognitive constraints formalized by a standard trial and error learning dynamics. Our findings support the claim, which we take to be independently attractive, that quality is a property of discourses among benevolent interlocutors whose interaction is constrained by the structure of assertion and rejection.

Some have argued that speakers are rationally required to be truthful (Grice, 1975; Lewis, 1972). Others have pointed out that the dutiful transfer of information from speaker to listener is too simplistic a picture of language use (Wilson & Sperber, 2002). From our perspective, these two observations can be squared. On the one hand, there are more equilibria than the purely truthful quality equilibrium in rejection games, as we have seen. Thus, more actions can be rational than truth-telling. On the other hand, plausible and cognitively modest assumptions lead to quality as the norm. Pragmatic regularities may be understood as (defeasible) descriptive generalizations about the typical behaviour of linguistic agents in contexts whose abstract features are formalized as particular kinds of games. A norm of truthfulness, what Grice (1975) dubbed the maxim of quality, is thus not an independent principle, but the natural order of a social setting determined by the interests of the interlocutors given assertion and rejection, and existing constraints such as sympathy.

Larry Horn (1984) has shown how to derive Grice's maxims, except for quality, from two general principles:

> The Q Principle (listener-based): Say as much as you can
> The R Principle (speaker-based): Say no more than you must

The Q and R principles are broad functional generalizations about the use of language. Horn's subsumption of the maxims under Q and R takes quality as a background assumption on which the application of the remaining maxims depends:

> The partial reductionist program I envision would retain the Maxim of Quality with its special character noted by Grice: unless Quality (or what Lewis, 1969 has called a Convention of Truthfulness) obtains, the entire conversational and implicatural apparatus collapses. (Horn, 1984, p. 12)

On this, Horn follows Grice:

> The observance of some of these maxims is a matter of less urgency than is the observance of others; a man who has expressed himself with undue prolixity would, in general, be open to milder comment than would a man who has said something he believes to be false. Indeed, it might be felt that the importance of at least the first maxim of Quality is such that it should not be included in a scheme of the kind I am constructing; other maxims come into operation only on the assumption that this maxim of Quality is satisfied. (Grice, 1975, p. 27)

> The maxim of Quality, enjoining the provision of contributions which are genuine rather than spurious (truthful rather than mendacious), does not seem to be just one among a number of recipes for producing contributions; it seems rather to spell out the difference between something's being, and (strictly speaking) failing to be, any kind of contribution at all. (Grice, 1989, p. 371)

With the present article, we have sought to complete the functionalist view of Gricean pragmatics by supplying an account of the emergence of quality. The rejection game shows that honest communication may result from the interaction between speaker and listener, provided we recognize that the listener has the means for rejection and the speaker is at least somewhat sympathetic.

Our proposed explanation may be summarized by the two following generalizations, which we take to be complementary to Horn's Q and R principles:

> The I Principle (listener-based): Reject inappropriate speech
> The A Principle (speaker-based): Avoid rejection

The I Principle encodes the Stalnakerian function of the speech act of rejection, as enacting a listener's veto on common ground management. The A Principle encodes the assumption that the speaker has an interest in making her own participation non-trivial. Rejection is a direct challenge to the speaker's overtly displayed goal of adding information to the common ground. Quality is a balance between these two drives.

ACKNOWLEDGEMENTS

We would like to thank audiences at the *Assertion and Proof* workshop (Università del Salento), the Dutch Research School in Philosophy (Amsterdam), and CogSci 2021 (Vienna). Special thanks to the reviewers and an editor for this journal. This work has received funding under the Horizon 2020 Program within the project *EXPRESS: From the Expression of Disagreement to New Foundations for Expressivist Semantics* (ERC Grant Agreement No. 758540), and the Horizon Europe Program within the project *Evolution of Logic* (MSCA Grant Agreement No. 101064835). Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

There are no data available.

ORCID


Giorgio Sbardolini https://orcid.org/0000-0003-0453-2445

REFERENCES

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115–1153.

Ahern, C., & Clark, R. (2017). Conflict, cheap talk, and Jespersen's cycle. *Semantics and Pragmatics*, *10*, 1–40.

Asher, N., & Lascarides, A. (2013). Strategic conversation. *Semantics and Pragmatics*, *6*, 1–62.

Barnett, S. A., Griffiths, T. L., & Hawkins, R. D. (2022). A pragmatic account of the weak evidence effect. *Open Mind*, *6*, 169–182.

Barrett, J. (2009). The evolution of coding in signaling games. *Theory and Decision*, *67*(2), 223–237.

Bicchieri, C. (2006). *The grammar of society*. Cambridge University Press.

Burge, T. (1975). On knowledge and convention. *Philosophical Review*, *84*(2), 249–255.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, *50*(6), 1431–1451.

Davidson, D. (1984). Communication and convention. *Synthese*, *59*(1), 3–17.

Davidson, D. (1986). A nice derangement of epitaphs. In E. Lepore (Ed.), *Truth and interpretation* (pp. 433–446). Blackwell.

De Jaegher, K., & van Rooij, R. (2014). Game-theoretic pragmatics under conflicting and common interests. *Erkenntnis*, *79*, 769–820.

Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, *10*(3), 103–118.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.
Frank, M. (2017). *Rational speech act models of pragmatic reasoning in reference games*. https://osf.io/x9mre/
Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, *4*, 1–81.
Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44. https://doi.org/10.1515/zfs-2016-0002
Frankfurt, H. G. (1986). *On bullshit*. Princeton University Press.
Goodman, N., & Frank, M. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.
Goodman, N., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.
Grice, H. P. (1975). Logic and conversation. In M. Ezcurdia & R. J. Stainton (Eds.), *The semantics-pragmatics boundary in philosophy* (pp. 47–59). Broadview Press.
Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.
Groenendijk, J., & Stokhof, M. (1984). *Studies on the semantics of questions and the pragmatics of answers* (Unpublished doctoral dissertation). University of Amsterdam.
Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, *3*(4), 367–388.
Hahn, U. (2014). The Bayesian boom: Good thing or bad? *Frontiers in Psychology*, *5*, 1–12.
Haldane, J. B. S. (1992). Animal communication and the origin of human language. *Current Science*, *63*(9–10), 604–611.
Hamblin, C. L. (1973). Questions in Montague English. *Foundations of Language*, *10*(1), 41–53.
Herrnstein, R. J. (1970). On the lawof effect. *Journal of the Experimental Analysis of Behavior*, *13*(2), 243–266.
Hollén, L., & Radford, A. (2009). The development of alarm call behaviour in mammals and birds. *Animal Behaviour*, *78*(4), 791–800.
Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schriffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11–42). Georgetown University Press.
Horn, L. (1989). *A natural history of negation*. University of Chicago Press.
Huttegger, S., & Zollman, K. (2011). Signaling games. In A. Benz, C. Ebert, G. Jäger, & R. van Rooij (Eds.), *Language, games, and evolution: Trends in current research on language and game theory* (pp. 160–176). Springer.
Incurvati, L., & Schlöder, J. J. (2017). Weak rejection. *Australasian Journal of Philosophy*, *95*(4), 741–760. https://doi.org/10.1080/00048402.2016.1277771
Kamenica, E. (2019). Bayesian persuasion and information design. *Annual Review of Economics*, *11*, 249–272.
Kamenica, E., & Gentzkow, M. (2011). Bayesian persuasion. *The American Economic Review*, *101*(6), 2590–2615.
Lachmann, M., Számadó, S., & Bergstrom, C. T. (2001). Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences*, *98*(23), 13189–13194.
Lederman, H. (2018). Uncommon knowledge. *Mind*, *127*(508), 1069–1105.
Lewis, D. (1969). *Convention*. Harvard University Press.
Lewis, D. (1972). Languages and language. In K. Gunderson (Ed.), *Minnesota studies in the philosophy of science* (pp. 3–35). University of Minnesota Press.
Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, *8*, 339–359.
Martinez, M., & Godfrey-Smith, P. (2016). Common interests and signaling games: A dynamic analysis. *Philosophy of Science*, *83*, 371–392.
Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge University Press.
Moran, P. (1962). *The statistical processes of evolutionary theory*. Clarendon Press.
Nash, J. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, *36*, 48–49.
Novak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, *364*, 56–58.
Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology* Retrieved from https://psycnet.apa.org/record/2022-91463-001, *152*, 346–362.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. MIT Press.
Penn, D. J., & Számadó, S. (2019). The handicap principle: How an erroneous hypothesis became a scientific principle. *Biological Reviews*, *95*(1), 267–290.
Pitchik, C., & Schotter, A. (1987). Honesty in a model of strategic information transmission. *The American Economic Review*, *77*(5), 1032–1036.
Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, *83*(5), 1281–1302.
Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*(6), 1–69.
Roth, A., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An experimental study. *The American Economic Review*, *81*(5), 1068–1095.
Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, *8*, 164–212.
Sally, D. (2000). A general theory of sympathy, mind-reading, and social interaction, with an application to the Prisoners' Dilemma. *Social Science Information*, *39*(4), 567–634.
Sally, D. (2003). Risky speech: Behavioral game theory and pragmatics. *Journal of Pragmatics*, *35*(8), 1223–1245.
Sbardolini, G. (2022). Is honesty rational? *The Philosophical Quarterly*, *72*(4), 979–1001.
Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.
Sobel, J. (1985). A theory of credibility. *The Review of Economic Studies*, *52*(4), 557–573.
Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Blackwell.
Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive Science*, *41*(3), 623–658.
Stalnaker, R. (1978). Assertion. *Syntax and Semantics*, *9*, 315–332.
Stalnaker, R. (1999). *Context and content*. Oxford University Press.
Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, *25*(5–6), 701–721.
Stalnaker, R. (2006). Saying and meaning, cheap talk and credibility. In A. Benz, G. Jäger, & R. van Rooij (Eds.), *Game theory and pragmatics* (pp. 83–100). Palgrave Macmillan.
Sumers, T. R., Hawkins, R. D., Ho, M. K., & Griffiths, T. L. (2021). Extending rational models of communication from beliefs to actions. *Proceedings for the 43rd annual meeting of the cognitive science society*, Virtual conference, 1–7. https://doi.org/10.48550/ARXIV.2105.11950
Tennant, N. (1999). Negation, absurdity and contrariety. In D. Gabbay & H. Wansing (Eds.), *What is negation?* (pp. 199–222). Springer.
van Rooij, R., & Sevenster, M. (2006). Different faces of risky speech. In A. Benz, G. Jäger, & R. van Rooij (Eds.), *Game theory and pragmatics* (pp. 153–175). Palgrave Macmillan.
Vignero, L. (2022). Updating on biased probabilistic testimony. *Erkenntnis*, 1–24. https://doi.org/10.1007/s10670-022-00545-7
Walker, M. A. (1996). Inferring acceptance and rejection in dialog by default rules of inference. *Language and Speech*, *39*(2–3), 265–304.
Wilson, D., & Sperber, D. (2002). Truthfulness and relevance. *Mind*, *111*(443), 583–632.
Zahavi, A. (1975). Mate selection—A selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214.
Zollman, K. J. S., Bergstrom, C. T., & Huttegger, S. M. (2013). Between cheap and costly signals: The evolution of partially honest communication. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1750), 1–8.

How to cite this article: Incurvati, L., & Sbardolini, G. (2023). The rejection game. *Mind & Language*, 1–22. https://doi.org/10.1111/mila.12460

APPENDIX: RESULTS OF SIMULATIONS

Experiments were written partly in NetLogo and partly in Python. The first set of experiments is about the interaction of rejection and sympathy. A simulation consists in 500 data points, each determined by a series of rounds of the rejection game. A series of rounds stops when the probability of success is higher than .999, or else after 10^5 rounds. We tested whether sympathy alone suffices for quality (with $\lambda = 1$), whether rejection alone does, and the effects of the two combined. In our simulations, the agents' chance of success with sympathy but no rejection in two states is .276, with rejection but no sympathy is .22. With sympathy and rejection combined, success is virtually guaranteed (close to 100%). Results are plotted in Figure 4.

In the course of a simulation, truthfulness increases as the speaker approximates the most optimal strategy of sending a true proposal in each state. Trust may be defined as the listener's acceptance of the speaker's proposal regardless of the state. Truthfulness and trust are typically correlated, but need not coincide: the speaker can induce the listener to trust her, by signalling truthfully at least in some states, but then exploit the acquired reputation by getting an assertion accepted in states in which it is false. In these cases, truthfulness drops while trust, rapidly accumulated at the beginning, remains high.

This effect tends to be more visible with more states, as Table A1 shows. For example, with five states, S selects TT with probability 34.23% (still remarkable, having to search through 5^5 strategies), but L accepts 79.92% of the signals. With six states, S selects TT with probability 5.36%, but L accepts 47.59% of the signals. These values appear to exceed what the speaker's behaviour would justify. Further work seems necessary to establish how robust this effect is, or whether it is due to computational limitations in our model.

Sympathy may not be the only way of achieving quality. In a second set of experiments, we studied two different heuristics, and a host of proposals based on them. Results are not as promising as in a sympathy-based account. One hypothesis is to introduce some variation to unsettle the speaker's learning inertia: speakers make random mistakes. To test this hypothesis, we ran simulations of a two-state game with random errors: occasionally, the speaker picks a signal regardless of past experience. Error frequency is $0 < \gamma \leq 1$. With $\gamma = 1$ an error might occur at every round of play with 1% chance.

Another idea is a version of Win-Stay/Lose-Randomize, a heuristics known to sometimes lead to optimal solutions in some two-player games (Novak & Sigmund, 1993). If the speaker succeeds she keeps doing what she does, but if she does not she changes her plan: mistakes affect behaviour just as successes do. In practice, when a message gets rejected in a given state, the probability that the same message will be sent again in that state gets a little lower, while

TABLE A1 Results of simulations for 100 trials with $\lambda = 1$.

No. of states	Random chance	% truthfulness	% trust
2	$1/2^2 = .25$	> .99	99.64
3	$1/3^3 = .037$	> .99	98.48
4	$1/4^4 = .003$	.95	95.11
5	$1/5^5 \approx 0$	.34	79.92
6	$1/6^6 \approx 0$	.05	47.59

TABLE A2 Likelihood of truthfulness (two alternatives).

	Result
Sympathy	$> .99$
Random-error, $\gamma = 0.2$	.19
Win-Stay/Lose-Randomize	.16
Random-error, $\gamma = 0.8$	.13
WS/LR + Random-error, $\gamma = 0.2$	.11
WS/LR + Random-error, $\gamma = 0.8$	.02

the probability that a randomly chosen message is sent in that state gets a little higher. For completeness, we also tested combinations of Random-error and Win-Stay/Lose-Randomize: after all, both are fairly intuitive and undemanding heuristics, supported by fairly natural assumptions. Results are reported below and are not encouraging.

One problem is that we do not know how variation induced by errors plays out with rejection. For example, if rejection makes the speaker less likely to send again the same message in a given state (relatively speaking), rejection itself becomes a disturbance. In general, however, although Random-error and Win-Stay/Lose-Randomize do not drive toward the best possible equilibrium, they do drive away from the worst, namely the community of liars equilibrium $\langle \mathrm{NN}, \mathrm{RR} \rangle$. Such an outcome is very unlikely under the two heuristics explored here and their combinations (Table A2).