Is Honesty Rational?*

Giorgio Sbardolini
ILLC and Philosophy Department, University of Amsterdam

August 29, 2022

Forthcoming in *The Philosophical Quarterly*. Please cite published version and do not circulate without permission.

Abstract

According to the Maxim of Quality, rational agents tend to speak honestly. Due to the influence of Grice, a connection between linguistic rationality and honesty is often taken for granted. However, the connection is not obvious: structural rationality in language use does not require honesty, any more than it requires dishonesty. In particular, Quality does not follow from the Cooperative Principle and structural rationality. But then what is honest rational speech? I propose to move the discussion to the context of Stalnaker's theory of assertion. From this perspective, although there is no most rational way to behave, Quality follows from the structure of Stalnakerian conversations if interlocutors are sensitive to credibility. In this case, honesty is built on the expectation of reciprocity, and it is an outcome and not a precondition of rational communication. A benefit of my discussion is that the account of linguistic rationality falls under the more general view of interactive rationality familiar from theoretical Economics.

Keywords: Pragmatics · Coordination · Rationality · Honesty · Cooperativity · Assertion · Rejection

1 The Truth and Nothing But The Truth

Linguists and philosophers in the tradition of H. Paul Grice contend that, when it comes to human communication, rationality requires speakers to be honest. Rational commu-

*Thanks to two anonymous reviewers for their feedback and encouragement. Special thanks to Craige Roberts and Luca Incurvati. This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 758540) within the project *EXPRESS: From the Expression of Disagreement to New Foundations for Expressivist Semantics*.

nication is typically understood as a cooperative enterprise, in which interlocutors share interests and goals, speakers tell the truth, and listeners trust them.

Some idealization is no doubt built into this widely held picture, but it is quite unclear what exactly is being idealized and why. Sometimes it is said that cooperativity makes for simple scenarios in which it is easy to track the interlocutors' actions, but of course a purely cooperative interaction is just as simple as a purely competitive one. Economists spend a great deal of time reflecting on rational action, but from the perspective of competition, not cooperation. Strikingly, a standard argument in economics is that, since there are no incentives to tell the truth in everyday conversations, speakers have equal chances of being honest and dishonest (Farrell and Rabin, 1996, 104). We seem to have two views of rational action, one from linguistics and one from economics, both involve some kind of idealization, but they teach us very different lessons about rationality and honest communication.

So is honesty rational? The answer, I think, is No. Rationality *per se* does not imply honesty, anymore than it implies dishonesty. I don't expect this answer to be particularly controversial, though this point is often overlooked. After some preliminaries in §2, I will contrast two Scenarios in §3 which make the Gricean connection between rationality and honesty appear problematic. The second and potentially more controversial part of the paper aims to uncover the conditions under which rationality does lead to honesty. I will start in §4 by introducing a Third Scenario and the notion of credibility. In §5, I will argue that honesty depends not on cooperativity but on the expectation of reciprocity, and that honesty may be rational if interlocutors are sensitive to credibility. This leads to a better understanding of the kind of rationality at work in a Gricean approach to language use.

My discussion of credibility is informed by Robert Stalnaker's (1999; 2014) influential theory of assertion, and ties in with how economists have been thinking about honesty, in particular through the work of Robert Axelrod (1984). In the end, we can reconcile what linguists and economists have to say about honesty, but this will require a more nuanced understanding of cooperativity in conversation. Contrary to conventional wisdom, honesty is an outcome and not a precondition of rational communication.

2 Background

Grice epitomized a requirement of honesty in his celebrated Maxim of Quality:

> Try to make your contribution one that is true (Grice, 1975, 46).

Despite the imperative mood, most linguists and philosophers read the Maxim as a description of how language is used that purports to be empirically adequate. This is the way pragmatic generalizations are understood in the Gricean tradition particularly after Robert Stalnaker: they maxims are, we might say, the natural laws of linguistic practice. Moreover, such laws are supported by our status as rational individuals: rather than *ad*

hoc stipulations they are supposed to follow from general principles that govern our social and mental life.

In this spirit, we should get rid of the imperative mood and of the voluntaristic 'try to'. A descriptive generalization inspired by the Maxim of Quality could be:

> Rational speakers tend to be truthful.

A couple of simple remarks. First, sometimes one tries to tell the truth, to the best of one's knowledge, but does not succeed. What we think we know is sometimes not what we know. Second, speakers are expected to tell the truth, if at all, relatively to the world they purport to be in for the purposes of the conversation—not the truth strictly and literally speaking. A conversation may be about make-believe, fictional, and counterfactual scenarios.

Let us set these simple remarks aside and frame the discussion in terms of honesty rather than truthfulness. By honesty I mean to knowingly say something one believes to be true relative to the world of interest, which may or may not be actual. Dishonesty is to knowingly say something one believes to be false relative to the world of interest. A Gricean connection between rationality and honesty could then be the following.

> **Quality**: Rational speakers tend to be honest.

Is Quality an adequate description of linguistic practice? Presumably, other attitudes besides honesty are necessary for rationality, and together sufficient, such as being relevant, being perspicuous, and so on. Yet one can't help but notice, with Wilson and Sperber (2002), that speakers fail to be honest a bit too often for us to complacently believe that, in ordinary conversations and odd exceptions aside, honesty rules. Politicians, salespeople, lobbyists, and poets, aren't always honest. Some speakers don't even try. There are liars, merely self-interested players, and Frankfurtian bullshitters (Frankfurt, 1986). Wilson and Sperber weigh in on examples of this kind to reject the whole edifice of Gricean pragmatics, and with it, the project of explaining pragmatic regularities in terms of rational choice (Sperber and Wilson, 1995).

I think that this is an overreaction. As I mentioned, it's not entirely clear what the Gricean assumptions are idealizing away, and therefore it is not clear how robust are the conclusions reached under the blanket of idealization. As correctly urged by Wilson and Sperber, a theory of linguistic communication ought to account for the speech of politicians, salespeople, and so on, on pain of empirical inadequacy. However, I also think that the project of explaining facts about language use by appeal to functional constraints derived from the theory of rationality is not threatened by this concession. After all, the question raised by Quality is not really empirical, but about the relation between Honesty and Reason itself. And yet, the idea that rational speakers are honest does ring a bit panglossian.

Before diving in, two more preliminary considerations. The first is about rationality. A distinction is often made between structural and substantial rationality (Kolodny,

2005; Williams, 2020). A structurally rational agent is, to a first approximation, someone (*i*) whose degrees of belief are consistent with probability theory, (*ii*) whose choices are consistent with decision theory, and (*iii*) who responds to evidence by Bayesian conditionalization. There may well be more conditions, and each of (*i*)-(*iii*) could be revised or refined, but in general structural rationality can be formally understood as overall coherence in means-end reasoning.

Structural rationality does not rule out the belief that the Democrats are pedophiles, or the desire for a saucer of mud. Yet there is a sense in which it is irrational to have such beliefs and desires. Substantial rationality is an account of the constraints we are subject to, relative to the particular contents which we believe and desire. Theories of substantial rationality are perhaps more familiar in meta-ethics than in linguistics, and it is not immediately obvious what a substantially rational linguistic agent would look like. There may well be a sound argument from some conception of substantial rationality to honesty—Kant famously tried one (Mahon, 2006). But the possibility of pragmatics as a descriptive theory of language use is not usually taken to depend on the soundness of such arguments. In contrast, we have a mathematically precise account of structural rationality, the details of which may be controversial but not its general shape. Henceforth I shall use 'rational' to mean structurally rational. The title applies primarily to individuals, and by extension to the actions they choose.

The final preliminary point is about the significance of Quality in a theory of communication. This point is about history and methodology, and it helps to circumscribe the limits of the present inquiry. Talk of honesty in conversation is ubiquitous in the philosophical and linguistic conventional wisdom. Why? As I mentioned, one idea is that honesty simplifies the discussion and helps to bring some distinctions into sharper focus. This may be true but it's not a particularly impressive consideration, and I suspect there are other reasons. Although this is not the place for an elaborate account of the historical evidence, I want to suggest that conventional wisdom has absorbed a certain foundational perspective on the use of language that Grice and others shared, including some explanatory worries about semantic content that come with this foundational perspective.

As is well known, Grice distinguishes natural meaning, on which smoke "means" fire, from non-natural meaning or 'meaning$_{NN}$':

> 'A meant$_{NN}$ something by x' is (roughly) equivalent to 'A intended the utterance of x to produce some effect in an audience by means of the recognition of this intention' (Grice, 1957, 385)

Suppose that we want to explain what the English sentence 'Snow is white' means. By Grice's meaning$_{NN}$, we could appeal to the fact that an utterance of 'Snow is white' typically has the effect of causing the audience to come to believe that snow is white. The audience will come to such belief upon hearing 'Snow is white' provided that the speaker's intention that the audience form such belief is properly recognized. Grice's account of the content of a sentence, therefore, presupposes that the speaker is honest—that is, that the

speaker has the "correct" intentions, which are then recognized by the audience. If the speaker isn't honest or isn't recognized by the audience as being honest, all sort of wrong mappings of utterances to beliefs would confuse Grice's account of content beyond repair.

Another important source of the idea that honesty plays a large role in language use may have been David Lewis (1969; 1972). Lewis's project is very different from that of Grice and his followers. Lewis's project turns in part on an explanation of what it is for a population to 'use' a language. As is well known, Lewis's explanation appeals to conventions of truthfulness and trust: (purported) behavioral regularities, according to which speakers tell the truth, and listeners believe them.

> the convention whereby a population P uses a language L is a convention of *truthfulness* and *trust* in L. To be truthful in L is to act in a certain way: to try never to utter any sentences of L that are not true in L. Thus it is to avoid uttering any sentence of L unless one believes it to be true in L. To be trusting in L is to form beliefs in a certain way: to impute truthfulness in L to others, an thus to tend to respond to another's utterance of any sentence of L by coming to believe that the uttered sentence is true in L. (Lewis, 1972, 167)

According to Lewis, the semantic content of a sentence can be read off its conventional use. This may be done relative to a particular convention in which, by and large, speakers say what they believe, and listeners believe what the speakers say. Lewis's account of content seems very different from Grice's, and both are controversial, but both depend on the assumption that language use is governed by honesty. Both have been very influential.

Lewis and Grice, and others after them, attempted to explain the intentionality of language (the fact that our utterances represent the world) in terms of the intentionality of belief and action. The general outline of the explanation appears to be, roughly: consider a context –call it *the context of origin*– and look at how language is used there. In the context of origin, we can trace the content of a sentence to what speakers believe, and what listeners thereby come to believe as an effect of its use. Quality comes in, in the context of origin, to guarantee the proper match between linguistic expressions and their contents. This allows us to reduce the semantic to the cognitive: for a sentence to have its content is, roughly, for it to be used as it is used in the context of origin.

An explanatory project of this kind may or may not succeed. However, even if it does succeed, there's no commitment to hold that the only context in which language is used rationally is the context of origin. Once linguistic expressions come to have a content, however this feat is accomplished, we can lift the pretense of supposing ourselves to be under a convention of truthfulness and trust, or that utterances mean something only if speakers have honest intentions.[1] The explanatory project of reducing meaning to belief

[1]Otherwise, the descriptive theory of language use would involve a commitment to utterly false assumptions. However, Grice does occasionally make some puzzling remarks.

> The maxim of Quality, enjoining the provision of contributions which are genuine rather than

and action is, needless to say, different from the pragmatic project, also largely inspired by Grice and Lewis (1979) among others, of describing language use in terms of general principles of rationality. Assumptions made for the former need not carry over to the latter.

3 Language Games

In this section I will introduce two Scenarios in which the same structural assumptions about the speaker's rationality lead to different conclusions about honesty. The point of this section is that cooperativity is a rather weak ground for Quality. In the next section I'll look for a better foundation. We begin with an Initial Case.

> *Initial Case*. Shiv has a job interview. The Royco Corporation is hiring in one of two positions: Chief Executive Officer (a_1) or office manager (a_2). Royco prefers that someone highly qualified is hired as CEO, and that someone less qualified is hired as office manager. Shiv is either highly qualified (t_1) or less qualified (t_2), and she knows which. Shiv can either say 'I'm highly qualified' (m_1), or 'I'm less qualified' (m_2). Rationality of both players is mutual knowledge, as well as Royco's preferences.

In the Initial Case there are two states, t_1 and t_2, one of which happens to be the case: Shiv is highly qualified or less qualified. She can send one of two messages to Royco: m_1 or m_2. This is a choice between honesty and dishonesty: she is honest if she says m_i if t_i is the case, and dishonest if she says m_i if t_j is the case $(j \neq i)$. Royco chooses between action a_1 and action a_2. The former is appropriate, from Royco's perspective, if Shiv is highly qualified, the latter if she is less qualified. Thus Royco's preferences are for being able to sort which of two possible states is actual. The roles of Shiv and Royco are those of speaker and listener respectively, and I will often refer to them as such in the following. I assume that both know to be rational, that the other is rational, and that both know the listener's preferences. I do not assume that the listener knows the speaker's preferences (which I haven't yet specified). The same assumptions also hold throughout the paper.

I also assume here and for the rest of the paper that talk is cheap: there are no special costs for telling the truth or lying. This is a natural assumption to make about ordinary conversations. However, it is theoretically important, for I am setting aside the costly signalling hypothesis. The costly signalling hypothesis is the base of a well known account of

> 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, 371)

This seems to imply that mendacious talk fails to be talk 'strictly speaking'. That's just bizarre, and I'm not sure what the 'strictly speaking' qualification is there for.

honesty in biology and economics (Zollman et al., 2013). Roughly, the idea is that costs are inflicted on honest behavior that cannot be balanced by the benefits of deception. So, only honest agents can afford to pay them.[2] For us, costs that support Quality are somewhat implausible: honest talk *per se* does not seem to require any special effort or waste.

The assumption that talk is cheap is shared with Lewis's well known signalling game (Lewis, 1969; Skyrms, 2010). In fact, there are many similarities between the Lewisian game and the interaction between Shiv and Royco. In both cases, the first player has some initial information, and sends a signal to the second player, who takes an action whose success depends on the initial information. But there are also some important differences, primarily due to the different goals of the discussion. It is unproblematic for Lewis to assume a preference for coordination, since his goal was to explain content in terms of the interaction. But it would be problematic for us to assume coordination in order to explain honesty, as I'll argue below. Conversely, it is unproblematic for us to assume that the agents use a common language in which they are already semantically competent: after all, this is how most conversations ordinarily are. Lewis could not make this assumption, for the obvious reason that his goal was to explain how semantic content came about. To put it briefly: Lewis assumed honesty (or rather: truthfulness) to explain content, I assume content to explain honesty.[3]

3.1 First Scenario

In the Initial Case, Shiv could tell Royco 'I'm highly qualified' if she is highly qualified, or 'I'm less qualified' if she is less qualified, but she could also lie. Will she be honest? Maybe! At the moment the Initial Case is too underspecified for an answer, since it does not include a description of Shiv's payoffs. With regards to this, we could just assume a preference for coordination right away. Thus we complete the description of the Initial Case by introducing a First Scenario:

> *First Scenario*. As in the Initial Case. In addition, if Shiv is highly qualified, she prefers to be CEO, because the office manager position would be boring. But if Shiv is less qualified, she prefers to be office manager, because working as CEO would be too stressful.

[2] In biology, this hypothesis surfaces as Zahavi's (1975) handicap principle (see also Lachmann et al., 2001); in economics, see Spence (1973).

[3] A further issue that would require attention concerns the notion of content at play here. Recent work on the Lewisian signalling game, after Skyrms (2010), builds on information-theoretic accounts of content (Dretske, 1981). For the purposes of this paper, it is fine to assume a qualitative notion of content, as in mainstream truth-conditional semantics and linguistic pragmatics. I am not sure that the two approaches to content are incompatible, but this issue definitely deserves a separate discussion. From an information-theoretic perspective, honesty is the topic of a rapidly growing literature. See Birch (2014); Shea et al. (2018); Fallis and Lewis (2019).

The First Scenario adds to the Initial Case the stipulation that Shiv prefers an outcome just in case Royco does, and Royco prefers a_i if t_i. So Shiv prefers a_i if t_i. The interaction between Shiv and Royco can thus be represented by Table 1. The players' preferences are perfectly aligned.

		R	
		a_1	a_2
S	t_1	$1,1$	$0,0$
	t_2	$0,0$	$1,1$

Table 1: Coordination

Coordination is one way for honest communication to come out of rational decision-making (Lewis, 1969; Rabin, 1990; Farrell and Rabin, 1996). For Shiv has no reason to lie now: since she prefers Royco to choose a_i if t_i, lying would only confuse Royco, increasing the chance of mistake. But Shiv doesn't want Royco to make a mistake. Therefore, she keeps honest.

A few more detail are useful for later. We zoom in on Shiv's choice between honesty and dishonesty. She knows that Royco is rational. In particular, she knows that Royco updates by Bayes Rule. Let's stipulate for simplicity that Royco chooses a_i if Royco's posterior belief that t_i is greater than $1/2$. By Bayes Rule, the posterior probability of t_i is given by Royco's prior weighted by the proportion between the conditional probability that message m is sent if t_i is the case and the unconditional probability that m is sent.

$$p'(t_i) = \frac{p(m|t_i) \times p(t_i)}{\sum_t p(m|t)} \qquad \text{Bayes Rule}$$

As we stipulated, Shiv prefers to be hired as CEO just in case she's highly qualified, and as office manager just in case she is less qualified. Given that Royco chooses a_i if $p'(t_i) > 1/2$, Shiv can deduce the probabilities of action that maximize the chance that her preferred outcomes obtain.[4]

$$\begin{array}{ll} p(m_1|t_1) = 1 & p(m_1|t_2) = 0 \\ p(m_2|t_1) = 0 & p(m_2|t_2) = 1 \end{array}$$

If Shiv approximates these probabilities, she tends to be honest. In addition, she behaves rationally relative to the other player: she maximizes expected utility while keeping fixed the assumption that her interlocutor is rational. Finally, it would be of no advantage

[4]Probabilities of action are conditional probabilities to send a message in a given state. Their metaphysical interpretation depends on how we interpret conditional probabilities, something on which I have nothing to say. If we are thinking about a single interaction, we can think of probabilities of action as propensities, dispositions, or tendencies for the speaker to act. Over a large number of interactions, we can think of them as frequencies distributed across space (many speakers) or time (one speaker multiple times).

for Shiv to deviate from these probabilities given how things are with Royco. Unilateral deviation would only increase the probability that the listener does not make the choice the speaker would prefer. Thus, these probabilities describe an equilibrium from the speaker's perspective. In technical terms, they describe an equilibrium of *the speaker's subgame*: the subset of the complete game obtained by considering the speaker's choice while holding fixed a rule for the listener's choice. The rule for the listener's choice is a deterministic one that depends on the posterior degree of belief: a_1 iff $p'(t_1) \geq 1/2$ and a_2 iff $p'(t_2) > 1/2$. The speaker's subgame is a natural description of the choice situation faced by the speaker, who thinks to herself: if the listener chooses an action rationally (if the listener updates by Bayes and chooses on the basis of posteriors) what should I rationally say? I will adopt the same perspective in the Second Scenario below.

The First Scenario shows how a speaker may be led to honesty on grounds of rationality. The reasoning is formally flawless, it seems to me, but not very fulfilling as an answer to the question whether honesty is rational. If we ask whether it's rational to be honest we would like to receive an explanation of the status of honesty. Assuming a preference for coordination only invites the question: and what explains the preference for coordination? The point is not one about empirical adequacy (as the issue raised by Wilson and Sperber (2002)), but about the value of such an explanation. For the explanation of honesty given in the First Scenario relies on a stipulation about preferences, and therefore it is only as good an explanation as stipulations are: not very good. The Initial Case tells us of a listener who prefers to make an informed decision about hiring someone, and the First Scenario describes a speaker who prefers that the listener makes an informed decision. But this is arbitrary: the speaker could prefer something completely different without thereby violating any constraint of structural rationality, as the Second Scenario shows.

3.2 Second Scenario

Assumptions about the rationality of linguistic agents can explain instances of language use even if coordination fails, and so presumably outside the purview of the Cooperative Principle (De Jaegher and van Rooij, 2014). To see this, consider a Second Scenario, also built on the Initial Case.

> *Second Scenario*. As in the Initial Case. In addition, Shiv prefers the CEO job unconditionally, because the pay is so much better.

Royco's preferences are the same as above, but now Shiv has an incentive to deceive. The players' playoffs in this case can be represented by Table 2. (In a pair of payoffs (x, y), x is the row player's payoff.)

What is the rational thing to do for Shiv? As above, I assume that Royco hires her as CEO just in case their posterior belief that t_1 is at least $1/2$, and I assume that Shiv knows that Royco is rational. She decides what to say accordingly. Clearly, if she is highly

		R	
		a_1	a_2
S	t_1	$1,1$	$0,0$
	t_2	$1,0$	$0,1$

Table 2: Partial Conflict

qualified, she will say that she is. Hence $p(m_1|t_1) = 1$ and $p(m_2|t_1) = 0$, as in the First Scenario. But what will she do if she's not highly qualified? Let's assume, for concreteness, that Royco's priors are $p(t_1) = 0.3$ and $p(t_2) = 0.7$, and that these priors are accurate. Without communication, Royco will hire Shiv as office manager since according to these priors she is more likely to be less qualified than highly qualified, and Royco has no further information.

With communication, Shiv may exploit the rationality of her interlocutor. Holding fixed that Royco integrates the information provided by Shiv with Bayes Rule, she can determine probabilities of action that maximize her chances to be hired as a CEO. This means solving for $p(m_1|t_2)$ in Bayes's equation so that $p'(t_1) = 1/2$.

$$\frac{1}{2} \times \sum_t p(m_1|t) = p(m_1|t_1) \times p(t_1)$$

Since $p(m_1|t_1) = 1$, Shiv should rationally lie if she is not qualified for the CEO job with probability $3/7$.

$$\begin{array}{ll} p(m_1|t_1) = 1 & p(m_1|t_2) = 3/7 \\ p(m_2|t_1) = 0 & p(m_2|t_2) = 4/7 \end{array}$$

Consider a garage sale of 100 items. The salespeople can advertise an item on sale as 'This is high price' (m_1) or 'This is low price' (m_2), and the items themselves are either high or low quality (t_1 and t_2). Potential customers are passing by. They expect that prices indicate quality, but unfortunately some low quality items might be advertised as high price and sold as such. Let's say that a potential customer buys if and only if they believe that a product is high quality with credence at least one half. We may suppose that the customers share a true prior that only 30% of the items on sale are high quality. If the salespeople signal m_1 if t_1 with probability 1, and m_1 if t_2 with probability $3/7$, then 60 items will be sold, namely the 30 that are high quality, plus 3 every 7 of those that are low quality. This is so even if the potential customers know that only 30 items are worth buying.

This outcome is rational for both speaker and listener. Since the listener is rational, sharing information by uttering m_1 or m_2 might increase or decrease their degree of belief. In other words, the listener is responsive to evidence and the utterance counts as evidence. The speaker can exploit this by finding the value of the probability of uttering m_1 or m_2

that maximizes her payoffs. Of course, the specific value depends on the priors, which in the example are arbitrary numbers. Whatever the value, however, the probabilities above define an equilibrium of the speaker's subgame.

Deviation leads the speaker to be worse off.[5] The speaker could lie with lower probability, but her expected payoff would decrease because the expected utility of dishonest behavior would decrease. Indeed, Shiv could even be perfectly honest, as in the First Scenario, but then her chances to be hired as CEO would be about 30%. Alternatively, the speaker could lie with greater probability, but her expected payoff would decrease because her message would be less informative and it is rational for Royco to disregard an uninformative message. Up to equilibrium point, it is rational for Royco to listen to what Shiv says, factoring it in as evidence for the hiring decision: after all, Shiv is $4/7$ likely to tell the truth in case she is less qualified, hence Royco does overall better listening to her than relying solely on the priors. But if Shiv lies more frequently, and if for example she just says 'I'm highly qualified!' in all cases, her message becomes uninformative, and so it would be rational (utility-maximizing) for Royco to revert to the priors rather than listening to it: it would be like deciding whether it's time to go by consulting a broken clock. Royco of course doesn't know Shiv's probabilities of action. But Royco is rational and Shiv knows it. So she reconstructs what is rational to do for her based on what is rational to do for Royco. Shiv's probabilities optimally balance the informativity of the message with the high expected utility of dishonesty.[6]

Of course, the lie need not be shameless. It might be just enough to present things as marginally better than honesty would strictly speaking prescribe. (Maybe Shiv is just embellishing her CV a bit aggressively. Economists euphemistically speak of 'information design.') In this regard, it is important to appreciate that there are constraints imposed by structural rationality on how much dishonesty is viable, since (*i*) it is only the outcome that the listener does not prefer that is affected by dishonesty, and (*ii*) the speaker should not lie so much that her message is no longer informative. This is a connection between rationality and honesty that should not be neglected. Still, the speaker does rationally lie.

There is an important difference of perspective between reasoning about a subgame as I did and reasoning about the game as a whole. Equilibria for the whole game of Table 2 may be found by dominance reasoning. If S is t_1, then R should choose a_1, and if S is t_2 then R should choose a_2. Conversely, if R chooses a_1 then S can go either way, likewise if R chooses a_2. Hence the game in the Second Scenario has two Nash equilibria (in pure strategies): (t_1, a_1) and (t_2, a_2). Dominance reasoning is a standard way of thinking about games, but it is a bit odd in the present context. For it is assumed that players have

[5] Thanks to an anonymous referee for pressing me to clarify this point.

[6] Both pure honesty are pure dishonesty can be in equilibrium in the Second Scenario. The first would have the same strategies as the familiar coordination case discussed above; the second would be what's known in game theory as the 'babbling equilibrium', in which the signal is ignored and rational listener's choice is determined uniquely by the priors. The Second Scenario builds on work on persuasion in economics (Pitchik and Schotter, 1987; Kamenica and Gentzkow, 2011; Kamenica, 2019).

strategies, and that they sift through them by eliminating those that are not as good as others. This assumption appears to make sense for games in which moves are made simultaneously, but not here, because it doesn't make a lot of sense to imagine that speakers choose what to say based on the assumption that the listener has already committed to choosing a_1 or a_2. In conversation, the speaker acts first and tries to influence the listener's decision. The speaker does not choose what to say on the basis of what the listener chose to believe. To capture this idea, it may be instructive to look at the speaker's subgame. Then, rather than assuming that the listener contemplates various stochastic maps from signals to actions and selects the best one by dominance reasoning, we assume that the listener treats the utterance as evidence for a deterministic decision. Then, holding fixed that this is how the listener rationally behaves, we reason backwards reconstructing what the speaker should do for her optimal benefit.

From this perspective, the same properties that apply to an honest speaker in the First Scenario apply to a partially dishonest speaker in the Second Scenario. In both cases, the speaker maximizes expected utility holding fixed the same deterministic rule for the listener's choice. Moreover, the strategies of honest and partially dishonest speaker are equilibria of the speaker's subgame relative to the two scenarios. If the honest speaker is rational in the First Scenario, the partially dishonest speaker is rational in the second. Of course, preferences differ in the two scenarios, but preferences are arbitrary.

Sometimes it is claimed that Quality follows from the Cooperative Principle along with the other maxims. However, it is apparent that the Cooperative Principle does not entail Quality without additional assumptions.

> The Cooperative Principle: Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. (Grice, 1975, 45)

The Second Scenario shows that by merely assuming structural rationality one does not get Quality out of the Cooperative Principle. In the Second Scenario Shiv does make the contribution that is required by the purposes for the talk she is engaged in: she abides by the letter of the Cooperative Principle. Perhaps the additional assumptions needed to get Quality out of rationality and the Cooperative Principle amount to establish that the purpose of the talk is to benefit all interlocutors, who then come to have the same preferences about the outcomes. But this just means that preferences are as in the First Scenario. And this, as I said, leads to a rather disappointing argument for Quality. For if honesty is rational because of a preference for coordination, then honesty is not so rational after all: there is nothing especially rational about preferring this or that. After all, 'It is not contrary to reason to prefer the destruction of the whole world to the scratching of my finger' (Hume, 1978, bk. 2, pt. 3).

We should trace honesty in conversation not to a stipulation about preferences, but to the interaction itself. We then let go of cooperativity as a precondition of rational communication, but we may find it again as a consequence of it.

4 A Third Scenario and Credibility

Is honest communication possible without a preference for coordination? There are some options (for a discussion, see (Skyrms, 2010, ch. 6)). One is the costly signalling hypothesis I mentioned above. Another option is to embrace uninformative messages (Crawford and Sobel, 1982; Martinez and Godfrey-Smith, 2016). Intuitively, if Shiv is not willing to reveal to Royco that she is less qualified, she could say 'I'm either highly qualified or less qualified', or something along these lines. This is honest, strictly speaking, but not informative. A third option is to assume that the speaker has some amount of sympathy or benevolence for the listener (Sally, 2000, 2003; Bicchieri, 2006), so that rather than different costs, there are different benefits attached to the speaker's choice. Here I will not follow these suggestions, as I would rather keep the payoff structure as it is, for generality, and communication informative. I will instead introduce a Third Scenario that formalizes very general aspects of conversational interactions. The Third Scenario expands upon the Second and will allow us to focus on the notion of credibility.

In the Second Scenario, Shiv has an interest in Royco believing that she is highly qualified. A natural reaction to this is to say that Royco should not listen to Shiv at all. One way to put the point is to say that Shiv's message 'I'm highly qualified' in the Second Scenario is not credible (Rabin, 1990; Sobel, 1985), because she would rather have her interlocutor believe that she is highly qualified no matter whether she is. For a preliminary definition, consider the following (cf. Stalnaker, 2006, 92-93):

> A message m is credible if, and only if, a rational speaker prefers that the listener believes m just in case m is true (relative to the world of interest in the conversation).

The preliminary definition can be revised or made more precise in various ways. In particular, it seems entirely plausible to say that a message can be more or less credible (there are degrees of credibility), that speakers themselves may be credible, and that whether someone or something is credible depends on the context and the topic of conversation. To keep it simple, however, I shall set aside these complications and treat credibility as a binary notion.[7]

Shiv is a rational speaker. However, in the Second Scenario she does not prefer that the listener believes her to be highly qualified just in case she is highly qualified. That's because she would also prefer Royco to believe that she is highly qualified if she is not. Hence, her message is not credible. Not knowing the speaker's preferences or probabilities of action, a listener should believe a message if it is credible, but not if it is not credible. Of course, if a message is not credible, it doesn't follow that the listener should believe something else, for a non-credible message could still be true.

[7] A notion of credibility plays a role in some debates on epistemic injustice (Fricker, 2007). I do not explore this apparent connection either.

What should a rational listener do, when the speaker's utterance lacks credibility? An intuitively plausible reaction to lack of credibility is dismissal. However, in the Second Scenario Royco doesn't really have the option to do so. If Shiv's claim to be highly qualified lacks credibility, the best Royco can do is to hire her as office manager. The Third Scenario accommodates the possibility of rejecting Shiv's utterance altogether.

> *Third Scenario*. As in the Second Scenario, except Royco has the option (a_3) to send Shiv home without hiring her at all. Royco prefers to hire rather than not, and prefers a highly qualified person as CEO and a less qualified person as office manager, but prefers not hiring to hiring the wrong person. Shiv prefers to be hired rather than not, and prefers to be hired as CEO rather than office manager.

In the Third Scenario, Royco has three options: to hire Shiv as CEO (a_1), to hire her as office manager (a_2), and to send her away (a_3). As above, Shiv is either highly qualified (t_1) or less qualified (t_2), and can either say that she's highly qualified (m_1) or that she is less qualified (m_2). Shiv's preferences are determined by her expected paycheck: to be hired as CEO beats to be hired as office manager, which beats unemployment. Royco prefers hiring the right person to not hiring, but prefers not hiring to hiring the wrong person. Table 3 reports these payoffs.[8]

		R		
		a_1	a_2	a_3
S	t_1	$2,2$	$1,0$	$0,1$
	t_2	$2,0$	$1,2$	$0,1$

Table 3: Partial Conflict with rejection

The payoff structure of Table 3 can also be interpreted as formalizing a very general and abstract description of the Stalnakerian dynamics of assertion. On Robert Stalnaker's (1978; 1999) influential account, conversations take place against a context defined by the presuppositions shared by the interlocutors. On this view, an assertion is

> a proposal to change the context by adding the content [of the utterance] to the information presupposed (Stalnaker, 1999, 10)

In all cases discussed in this paper, Shiv can make one of two assertions, so that she has the option to be honest or not. This is very natural, since there are always many things we

[8]The Third Scenario builds on previous work, especially by Rabin (1990), who describes a similar game briefly discussed by Stalnaker (2006) too. I ignore, for simplicity, at least one element that Stalnaker emphasizes, namely the distinction between the credibility of a message, and the belief that a message is credible. A message might be credible while nobody believes that it is. This distinction is no doubt important, but introduces complications that need not be dealt with at present.

can assert, and some big or small departure from what is strictly speaking true is likely to be always, or almost always, an option. Royco's actions a_1 and a_2 can be interpreted as the actions of coming to believe that the job candidate is highly qualified or less qualified, and hiring accordingly. These are the actions of coming to believe what Shiv says when she asserts to be highly qualified, or less qualified. By choosing a_1 or a_2, Royco proceeds to do their part in changing the context according to Shiv's assertion. Such context change is 'the essential effect of an assertion' (Stalnaker, 1978, 86). But there is always a "walk away" option.

> [T]he essential effect of an assertion ...is avoided only if the assertion is rejected (Stalnaker, 1978, 86)

The listener can refuse to accept what the speaker says. This can be for many reasons, lack of credibility being one of them. Rejection is action a_3, and in the Third Scenario we have a listener who can reject a job candidate whose assertions are not credible. But the speaker does not want her assertions to be rejected, for this would frustrate the purpose of making them. So the listener in the Third Scenario can retaliate against dishonesty. I will refer to interactions with this structure as Stalnakerian conversations.

In a Stalnakerian conversation, credibility can be a reference point for sorting equilibria. Suppose that Shiv sends m_1. Royco may reason as follows: m_1 is either credible or not. If it is credible, Shiv will send it only if t_1 is the case. Hence, Shiv doesn't send it if t_2. Consequently, m_2 is also credible, because Royco knows that if Shiv meant Royco to believe t_1, she would send m_1, not m_2. Therefore, if m_1 is credible, so is m_2, and the game will be resolved as a game of coordination. Suppose instead m_1 is not credible. Then Shiv would prefer to send m_1 even if t_2 is the case. So if Royco receives m_1, Royco should not believe that t_1 is the case and play a_1, but rather play a_3 and walk away. Should Royco believe at least m_2? No, because Shiv knows that Royco would not believe m_1, for m_1 is not credible. Since Shiv prefers that Royco plays a_2 over a_3, then m_2 is not credible either. Therefore, if m_1 is not credible, neither message is credible. Consequently, Royco should play a_3 no matter what.[9]

Two interesting equilibria stand out: one in which the message is credible, the speaker is always honest, and the listener never rejects; another in which the message is not credible, and the listener always rejects. The latter outcome is perhaps not particularly desirable for either, but neither can improve their payoff given what the other does. A general

[9]This conclusion depends on the stipulation made in the text that the listener should reject messages that lack credibility. This assumption seems fine for present purposes, but not very nuanced. As pointed out by an anonymous referee, there is also an equilibrium of the Third Scenario in which Shiv is partially dishonest but Royco hires her as office manager: Royco may still be better off doing that so long as Shiv does not lie too much. This is so even though Shiv's message is not credibile by the definition given. As the referee suggests, it would be perhaps more adequate, especially in the more general context of mixed strategies, to revise the definition of credibility given above and treat it as a gradable notion. Then we would have partially credible messages and we could specify a more nuanced rule of rejection. I would like to thank the reviewer for pressing this point.

lesson applies to Stalnakerian conversations. Even if the speaker's preferences or her probabilities of action are unknown to the listener, honesty in communication varies with the credibility of the message.

5 When Honesty Is Rational

Time matters, for different strategies play out differently over time. What is best to do for the speaker depends not only on what the listener does –this is always true in any non-trivial game– but also on what the listener will do given what the speaker previously did, and *vice versa*. Shiv has only one job interview with Royco, but it is likely that she will have to deal with Royco in the future too. The second time they interact it will not be the same job interview, or it will not be a job interview at all, but insofar as the structure of Table 3 is common to the Stalnakerian dynamic of assertion and rejection, the same choices will be available. Therefore, since they might meet again, Shiv might not want to put Royco in a position to retaliate against her dishonesty. Time affects Royco's choice as well, for if Shiv lied today, why believe her tomorrow?

Time changes the equation in the Third Scenario more than it does in the First and Second Scenarios, because the Third Scenario includes a discounting factor for the speaker's choice: the possibility of future rejection against a dishonest assertion. In this section, I will develop an account of reciprocity-based conversation, inspired by Robert Axelrod (1984) and subsequent work in economics, by bringing together honesty and credibility. On the general view I will sketch, linguistic rationality at work in language use can be reconciled with economic rationality, and I will focus on some of the assumptions that make such analysis possible.[10]

As soon as we allow for the possibility of an iterated game, a large number of strategies become available for each player, which depend on what the other player did the previous round. In order to keep the discussion manageable, I will only discuss a few carefully chosen strategies to make a general point. Let's assume that Royco can play ALWAYS ACCEPT, ALWAYS RETALIATE, or TIT-FOR-TAT. With ALWAYS ACCEPT, Royco always accepts Shiv's assertion, whatever she says. If she says she is highly qualified, Royco hires her as CEO; if she says she is less qualified, Royco hires her as office manager. ALWAYS ACCEPT exposes Royco to exploitation, for Royco never resorts to rejection. With ALWAYS RETALIATE, Royco is willing to accept what Shiv says on the first round, and continues to do so for as long as she speaks honestly, but as soon as she fails to be honest, Royco rejects thereafter. Royco starts nicely with TIT-FOR-TAT as well, and comes to believe what Shiv says on the first round. But then, Royco reciprocates whatever Shiv did the previ-

[10]My discussion in this section is based on well-established results in economic theory, sometimes known as "folk theorems" of repeated games. An early reference is (Luce and Raiffa, 1957, 102), and seminal results were due to Friedman (1971); Aumann and Shapley (1994), and Rubinstein (1994). The mathematical details had been worked out most extensively by Robert Aumann (1959, 1960, 1961).

ous round: if she has been honest, he accepts her assertion, but if she has been dishonest, he rejects. In TIT-FOR-TAT, Royco is quick to punish dishonesty and quick to forgive. In ALWAYS RETALIATE, Royco is quick to punish, and thoroughly unforgiving. For Shiv, the possibility that Royco is playing something other than ALWAYS ACCEPT changes things considerably. She might get advantage of Royco in one round, but then if Royco retaliates she would get her worst payoff, and might regret not being honest in that previous occasion.

As to Shiv, let's assume that she can play DISHONEST, HONEST, or DEVIANT TIT-FOR-TAT. With the first, Shiv is partially dishonest, for she always says 'I'm highly qualified' even if she is not. With the second, she speaks honestly: m_1 if t_1 and m_2 if t_2. With DEVIANT TIT-FOR-TAT, Shiv plays with the probabilities that maximize her payoff found in the Second Scenario (whatever the priors), and in addition she plays dishonestly if she was rejected at the previous round. DEVIANT TIT-FOR-TAT is the strategy of a smart but vengeful speaker. There are, of course, many other strategies for both of them.

I assume that with non-zero chance w, Shiv and Royco meet again after they have met once. Moreover, I assume that they remember what happened the previous time they met. From these two assumptions, it follows that the future casts a shadow over the present, which we can discuss in the form of cumulative payoffs for Shiv and Royco. For simplicity, the chance to meet again remains constant throughout. If $u(i, n)$ is the payoff for player i at round n, total payoffs for the interaction for player i are calculated as an infinite sum.

$$u(i) = u(i, 0) + w \cdot u(i, 1) + w^2 \cdot u(i, 2) + w^3 \cdot u(i, 3) + \ldots$$

Thus interactions that are very far into the future contribute little to current payoffs (for w^n tends to 0 as n tends to infinity), but still do. For illustration, suppose that a player receives the same payoff x at every turn, because both play a strategy that is constant over time. Then, since $1 + w + w^2 + \ldots = 1/(1 - w)$, the infinite sum reduces to $x/(1 - w)$.

$$u(i) = x + x \cdot w + x \cdot w^2 + x \cdot w^3 + \ldots = \frac{x}{1 - w}$$

Two such constant strategies are DISHONEST for Shiv and ALWAYS ACCEPT for Royco. In this combination, Shiv always gets a payoff of 2 (cf. Table 3). Suppose furthermore that the chance to meet again is $1/2$. Then the total payoff for Shiv in these favorable circumstances is 4. But suppose that Royco plays ALWAYS RETALIATE instead. Then Shiv gets 2 the first round, and forever 0 after the first round at which she falsely declared herself highly qualified—how soon this will be depending on the priors. For example, if she speaks dishonestly already at the first round, her total cumulative payoff will be 2. So the possibility of effective retaliation means trouble for a dishonest speaker.

Against TIT-FOR-TAT, a DISHONEST speaker does slightly better than against ALWAYS RETALIATE. As above, she receives a payoff of 2 after the first round, and then either 2 or 0 at each round depending on the chance that she was indeed highly qualified at the

previous round. But against TIT-FOR-TAT, it is possible for the speaker to do much better by playing HONEST. For then for every round at which a DISHONEST speaker gets 0 due to deception at the previous round, the HONEST speaker gets 1, which is the payoff for being hired as office manager when one presents oneself as less qualified. However, for every round at which a DISHONEST speaker gets 2, the HONEST speaker also gets 2. So honesty pays off against a listener who reciprocates by playing TIT-FOR-TAT.

HONEST is also no worse than DISHONEST against ALWAYS RETALIATE, for sooner or later the DISHONEST speaker will lie, receive 0 payoff and always 0 thereafter. In contrast, an HONEST speaker might score less on any single interaction, but never give the listener the chance to end the stream of positive payoffs. Of course, against an ALWAYS ACCEPT listener the HONEST speaker does not do as well as the DISHONEST speaker does, for the latter is better at exploiting gullibility.

Consider now strategies for the listener. ALWAYS ACCEPT performs worse than the other two options against DISHONEST, but TIT-FOR-TAT and ALWAYS RETALIATE are equivalent against HONEST. However, against a DEVIANT TIT-FOR-TAT speaker TIT-FOR-TAT does worse than ALWAYS ACCEPT. This is because as soon as the DEVIANT TIT-FOR-TAT speaker fails to tell the truth once, TIT-FOR-TAT retaliates with rejection, and then a cycle of rejections and dishonest assertions begins with poor payoffs for both. In contrast, a very forgiving ALWAYS ACCEPT listener doesn't fall into this trap.

From this survey we can extract a general conclusion. If the chance of future encounters is high enough, there is no uniquely best strategy independent of the strategy of the other player. Indeed, DEVIANT TIT-FOR-TAT is designed precisely to make this point by showing that HONEST is not uniquely best. Given any strategy, and a sufficiently high chance to play again, it is always possible to design an opponent against whom that strategy would do worse in the long run than an alternative strategy. On the other hand, if the chance of a future encounter is too low, so that the speaker has no reason to think that she will suffer negative consequences for her actions, nothing prevents her from rationally deceiving the listener. The general rule of economic rationality holds of rational speech act theory as well.

While there is no uniquely most rational behavior that makes one perform well against any opponent, HONEST and TIT-FOR-TAT stand out. These strategies are sensitive to the credibility of the message.

> A strategy is sensitive to credibility if, and only if, over time credible messages tend to be rewarded and non-credible messages tend to be discouraged.

HONEST is sensitive to credibility, since its success depends on how rewarding it is to send credible messages in the long run. In contrast, since the messages of a DISHONEST speaker are never credible, the success of DISHONEST over time does not depend on credibility, but only at best on finding a sufficiently credulous opponent. Moreover, TIT-FOR-TAT is sensitive to credibility. Indeed, TIT-FOR-TAT is designed to reward credible messages and

punish non-credible messages. In contrast, ALWAYS ACCEPT is not sensitive to credibility, and ALWAYS RETALIATE ceases to reward credible messages after the first breach of credibility. Strategies that are sensitive to credibility do best against strategies that are sensitive to credibility. However, strategies that are not sensitive to credibility might be much more successful against strategies that are also not sensitive to credibility. However, their success in this case depends on finding the "right" opponent. Therefore, although credibility-sensitive strategies need not lead to highest payoffs absolutely, their success depends on establishing solid grounds for lasting cooperation.

The argument in this section is inspired by Robert Axelrod's (1984) work on the iterated Prisoner's Dilemma (see also Kendall et al. (2007); Nowak and Sigmund, 1993). It is well known that in the one shot Prisoner's Dilemma it is rational for both players to defect (see Table 4). It is also well known that in the Iterated Prisoner's Dilemma it may be possible for an optimal equilibrium to emerge, particularly if players reciprocate on nice behavior. Indeed, some of the formal properties of reciprocity in the Iterated Prisoner's Dilemma hold as well in iterated Stalnakerian conversations. In particular, these are the possibility of effective retaliation, and the high-enough probability of a repeated interaction.

		B	
		Be Nice	Defect
A	Be Nice	$2, 2$	$0, 3$
	Defect	$3, 0$	$1, 1$

Table 4: The Prisoner's Dilemma

However, Stalnakerian conversations are not Prisoner's Dilemmas. The Prisoner's Dilemma has only one Nash equilibrium (in pure strategies), while the game of Table 3 has two: the message is credible and the listener always accepts, or the message is not credible and the listener always rejects (again, in pure strategies). So they are different games from a mathematical perspective.[11] From a philosophical perspective, the Prisoner's Dilemma does not seem to be a good model of conversation. Of course, some interactions are Prisoner's Dilemmas and some of them play out in language, but there is nothing about conversations as such, it seems to me, that should make us think that they display the conflict of interests that defines the Prisoner's Dilemma. Instead, the game I have focused on formalizes conversations as Stalnakerian interactions of assertion and rejection. Finally, there is also a difference of perspective. The literature on the Iterated Prisoner's Dilemma has often focused on the somewhat unproductive search for strategies that would outperform the currently best performing strategy (Binmore, 2015), but

[11] To some extent, there are similarities between Stalnakerian conversations and the Optional Prisoner's Dilemma in which the players can decide to not play (Batali and Kitcher, 1995; Orbell and Dawes, 1993). There are also similarities with the Stag Hunt (Skyrms, 2001), which has a unique Pareto optimal equilibrium, as the game in Table 2.

that's not my concern here. As we have seen, honesty does not necessarily lead to the highest cumulative payoffs.

My concern in this section has been to explore the connection between pragmatic and economic rationality, and to introduce some of the notions that bring out such connection. Insofar as generalizations that are true of economic reason apply also to the theory of language use, Gricean rationality is a species of economic rationality. I have argued that honesty in conversation can be supported by the possibility of rejection, and that this may be so because honesty over time is sensitive to credibility. Moreover, sensitivity to credibility is a watershed notion for conversational strategies, separating out those that lead to stable and better outcomes when combined with each other, from those whose success or failure depend on the opponent. In this sense, cooperativity may emerge among credibility-sensitive strategies, regardless of arbitrary preferences, and driven instead by structural features of the interaction.

6 Conclusion

Quality is the claim that rational speakers tend to be honest. In one shot interactions as the First and Second Scenario, the speaker's idiosyncratic preferences determine whether or not honesty is rational. So honesty is not rational—not without further qualification. With repeated interactions, and under a better approximation of natural conversations along the lines of Stalnaker's account of assertion, honesty sometimes maximizes the speaker's payoff but not against any kind of listener. Honesty does well in terms of credibility, but credibility-sensitive strategies are not always the best against strategies that are not sensitive to credibility, and so they are not necessarily the utility maximizing ones. However, credibility emerges as a natural boundary for conversational strategies, and honesty falls among those for which cooperativity can be an equilibrium over time.

For this conclusion, I have assumed that there is a sufficiently high chance that speaker and listener meet again after having met any given time, and that they have memory of their previous interaction. Moreover, I have assumed that it should be possible for the listener to retaliate effectively against a dishonest speaker, and that the speaker has the choice of honest or dishonest talk. We now have the beginning of an explanation of honesty in conversation that goes beyond the stipulation that people prefer to coordinate. Honesty is an outcome, not a precondition of communication, so long as conversations repeated over time have the structure of Stalnakerian assertion and rejection. Interlocutors may prefer to coordinate or not, but Quality depends on the proper management of credibility. If interlocutors are sensitive to the credibility of their messages, Quality emerges.

References

Aumann, R. (1959). Acceptable points in general cooperative n-person games. In *Contributions to the Theory of Games*, pp. 287–324. Princeton, NJ: Princeton University Press.

Aumann, R. (1960). Acceptable points in games of perfect information. *Pacific Journal of Mathematics 10*, 381–417.

Aumann, R. (1961). The core of a cooperative game without side payments. *Transactions of the American Mathematical Society 98*, 539–539.

Aumann, R. and L. S. Shapley (1994). Long-term competition: A game-theoretic analysis. In *Essays in Game Theory*, pp. 1–15. New York, NY: Springer.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York, NY: Basic Books.

Batali, J. and P. Kitcher (1995). Evolution of Altruism in Optional and Compulsory Games. *Journal of Theoretical Biology 178*, 161–171.

Bicchieri, C. (2006). *The Grammar of Society*. Cambridge: Cambridge University Press.

Binmore, K. (2015). Why all the fuss? the many aspects of the prisoner's dilemma. In M. Peterson (Ed.), *The Prisoner's Dilemma*, pp. 16–34. cambridge University Press.

Birch, J. (2014). Propositional content in signalling systems. *Philosophical Studies 171*, 493–512.

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

De Jaegher, K. and R. van Rooij (2014). Game-Theoretic Pragmatics under Conflicting and Common Interests. *Erkenntnis 79*, 769–820.

Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Press.

Fallis, D. and P. J. Lewis (2019). Toward a formal analysis of deceptive signaling. *Synthese 6*, 2279–2303.

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

Frankfurt, H. G. (1986). *On Bullshit*. Princeton University Press.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Friedman, J. W. (1971). A non-cooperative equilibrium for supergames. *The Review of Economic Studies 38*, 1–12.

Grice, H. P. (1957). Meaning. *Philosophical Review 66*, 377–388.

Grice, H. P. (1975). Logic and conversation. In M. Ezcurdia and 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*. Cambridge, MA: Harvard University Press.

Hume, D. (1978). *A Treatise of Human Nature*. Oxford: Oxford University Press.

Kamenica, E. (2019). Bayesian persuasion and information design. *Annual Review of Economics 11*, 249–272.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*, 2590–2615.

Kendall, G., X. Yao, and S. Y. Chong (2007). *The iterated prisoners' dilemma: 20 years on*, Volume 4. World Scientific.

Kolodny, N. (2005). Why be rational. *Mind 114*(455), 509–563.

Lachmann, M., S. Számadó, and C. T. Bergstrom (2001). Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences 98*, 13189–13194.

Lewis, D. (1969). *Convention*. Cambridge MA: 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.

Luce, D. and H. Raiffa (1957). *Games and Decision*. New York: John Wiley.

Mahon, J. E. (2006). Kant and the perfect duty to others not to lie. *British Journal for the History of Philosophy 14*, 653–685.

Martinez, M. and P. Godfrey-Smith (2016). Common Interests and Signaling Games: A Dynamic Analysis. *Philosophy of Science 83*, 371–392.

Nowak, M. and K. Sigmund (1993). A strategy of Win-Stay, Lose-Shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature 364*, 56–58.

Orbell, J. and R. Dawes (1993). Social Welfare, Cooperators' Advantage and the Option of Not Playing the Game. *American Sociological Review 58*, 787–800.

Pitchik, C. and A. Schotter (1987). Honesty in a model of strategic information transmission. *American Economic Review 77*, 1032–1036.

Rabin, M. (1990). Communication between rational agents. *Journal of Economic Theory 51*, 144–170.

Rubinstein, A. (1994). Equilibrium in supergames. In *Essays in Game Theory*, pp. 17–27. New York, NY: Springer.

Sally, D. (2000). A General Theory of Sympathy, Mind-Reading, and Social Interaction, with an Application to the Prisoners' Dilemma. *Social Science Information 39*, 567–634.

Sally, D. (2003). Risky speech: behavioral game theory and pragmatics. *Journal of Pragmatics 35*, 1223–1245.

Shea, N., P. Godfrey-Smith, and R. Cao (2018). Content in simple signalling systems. *British Journal for the Philosophy of Science 69*, 1009–1035.

Skyrms, B. (2001). The stag hunt. *Proceedings and Addresses of the American Philosophical Association 75*, 31–41.

Skyrms, B. (2010). *Signals: Evolution, Learning, and Information*. New York: Oxford University Press.

Sobel, J. (1985). A theory of credibility. *The Review of Economic Studies 52*, 557–573.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics 87*, 355–374.

Sperber, D. and D. Wilson (1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Stalnaker, R. (1978). Assertion. *Syntax and Semantics 9*, 315–332.

Stalnaker, R. (1999). *Context and Content*. Oxford: Oxford University Press.

Stalnaker, R. (2006). Saying and meaning, cheap talk and credibility. In A. Benz, G. Jäger, and R. van Rooij (Eds.), *Game Theory and Pragmatics*, pp. 83–100. London: Palgrave Macmillan.

Stalnaker, R. (2014). *Context*. Oxford University Press.

Williams, J. R. G. (2020). *The Metaphysics of Representation*. Oxford University Press.

Wilson, D. and D. Sperber (2002). Truthfulness and Relevance. *Mind 111*, 583–632.

Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology 53*, 205–214.

Zollman, K. J. S., C. T. Bergstrom, and S. M. Huttegger (2013). Between cheap and costly signals: the evolution of partially honest communication. *Proceedings of the Royal Society B: Biological Sciences 280*, 1–8.