Argumentation with (Bounded) Rational Agents

Robert van Rooij and Kris de Jaegher

Abstract A major reason for our communication is to influence our conversational partners. This is so both if our preferences are aligned, and when they are not. In the latter case, our communicative acts are meant to manipulate our partners. We all know that attempts to manipulate are nothing out of the ordinary. Unfortunately, the standard theory of rational communicative behavior predicts that any such attempt will be seen through and is thus useless. The main aim of this chapter is to investigate which assumptions of the standard theory we have to give up to account for our communicative behavior, when preferences between partners are not aligned.

1 Introduction: Communicate to Influence Others

Why do we talk? What is the purpose of our use of language? It is obvious that language is used for more than one purpose. Sometimes we use language with an *expressive* purpose: guess what our roommate just did when his computer crashed again. Sometimes language is being used to *strengthen relationships* between people: Our colleagues gossip a lot during lunch. We have to admit, however, that we normally use language just to *influence* the *behavior* of others. And to be honest, we think you are exactly like us. Indeed, although language is a multipurpose instrument, the purpose to influence other's behavior seems to be basic.

We would like to thank Frank Zenker for valuable comments and corrections and the participants of the Bayesian argumentation workshop for useful discussion.

R. van Rooij (✉)
Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands
e-mail: R.A.M.vanRooij@uva.nl

K. de Jaegher
Utrecht School of Economics, Utrecht University, Utrecht, The Netherlands

F. Zenker (ed.), *Bayesian Argumentation: The practical side of probability*, Synthese Library 362, DOI 10.1007/978-94-007-5357-0_8,

Table 1

	a_1	a_2
t_1	1,−2	0,0
t_2	1,3	0,0
t_3	1,−2	0,0

Now, why do we want to influence each others behavior, and how are we going to analyze this? Well, let us again speak for ourselves: We want to influence your behavior by our use of language because we believe that your changing behavior would be profitable for *us*. So we consider one communicative act better than another, when we expect the former to have a more profitable effect than the latter. This suggests that language use is very much like other kinds of economic behavior and that it can be studied profitably by means of *decision* and *game theory*.

According to decision theory, an agent should choose that action which has the highest expected utility. Consider now the following decision problem with an agent wondering which of $\{a_1, a_2\}$ she should perform:

	a_1	a_2
t_1	−2	0
t_2	3	0
t_3	−2	0

On the assumption that the three states are equally likely, it is clear that the agent will choose action a_2 because that has, on average, a higher utility than action a_1, 0 versus –1/3. If an agent receives new information, and the agent believes it, this will turn the old decision problem into a new one. Suppose, for instance, that the agent receives the information that the actual state is in $\{t_1, t_2\}$. As a reaction, the agent will adapt her probability function such that the posterior function gives a probability 0 to state t_3. Maximizing expected utility with respect to this new probability function now results in action a_1 because that has now, on average, the highest utility: 1/2 versus 0. But now suppose that the actual state is actually t_1. Although on the basis of this new information it was *rational* for the agent to choose for a_1, it still was *actually* the *wrong decision*.

Until now we assumed that our agent simply received truthful information. We haven't considered how she received it. Suppose that she received it from another agent. This other agent might also care about which action our agent is going to perform. For instance, he might prefer our agent to perform a_1 instead of a_2, independent of which state actually holds. In such a case, the combined utility table of the answerer (first entry) and our agent can be pictured as follows (Table 1).

For a situation that can be modeled by the above multiagent decision table, it makes a lot of sense for the informer to provide our agent with information $\{t_1, t_2\}$ in case the actual state is t_1: If the agent just accepts what she is informed of, she will perform action a_1, which is exactly what the informer hoped for. Thus, if our agent takes the new information at face value, she can be *manipulated* by the informer and will act in accordance with his, but not her own, preferences.

Table 2

	t_1	t_2		m_1	m_2		ρ_1	ρ_2	ρ_3	ρ_4
σ_1	m_1	m_1	ρ_1	a_1	a_1	σ_1	x, x	x, x	y, y	y, y
σ_2	m_1	m_2	ρ_2	a_1	a_2	σ_2	x, x	$1, 1$	$0, 0$	y, y
σ_3	m_2	m_1	ρ_3	a_2	a_1	σ_3	x, x	$0, 0$	$1, 1$	y, y
σ_4	m_2	m_2	ρ_4	a_2	a_2	σ_4	x, x	y, y	x, x	y, y

But now suppose that our agent knows the preferences of the informer as well and that this, in fact, is common knowledge. If she is also rational, our agent will see through the attempt of manipulation of the informer and will not take the new information at face value. If the informer is rational as well, she will see trough this in turn and will realize that it doesn't make sense to provide information $\{t_1, t_2\}$, because the acting agent won't take this information to be credible. A new question comes up: How much can an agent credibly communicate in a situation like that above? This type of question is studied by economists making use of signaling games.

2 Signaling Games and Nonaligned Preferences

In his classic work on conventions, Lewis (1969) proposed to study communication by means of so-called signaling games. In this section, we will only consider cheap talk games: games where the messages are not directly payoff relevant. A signaling game with payoff irrelevant messages is a sequential game of incomplete information with two players involved, player 1, the sender, and player 2, the receiver. Both players are in a particular state, an element of some set T. Player 1 can observe the true state, but player 2 cannot. The latter has, however, beliefs about what the true state is, and it is common knowledge between the players that this belief is represented by probability function P over T. Then, player 1 observes the true state t and chooses a message m from some set M. After player 2 observes m (but not t), he chooses some action a from a set A, which ends the game. The utilities of both players are given by $U_1(t, a)$ and $U_2(t, a)$. The (pure) strategies of the player 1 and player 2 are elements of $[T \rightarrow M]$ and $[M \rightarrow A]$, respectively. In simple communication games, we call these functions sending and receiving strategies, that is, σ and ρ.

What strategy combinations are *equilibria* of the game depends on the probability distribution. With distribution P, the strategy pair $\langle S, R\rangle$ is an equilibrium if, as usual, neither player can do any better in terms of expected utility by unilateral deviation. As a small example, consider the signaling game with only two states t_1 and t_2, two messages m_1 and m_2, and two actions a_1 and a_2, and utility functions $U(t_i, a_j) = 1$, if i = j, o otherwise. Obviously, both players have four (pure) strategies each. Furthermore, let $x = P(t_1) > P(t_2) = y$. Then, we have the payoff matrix in Table 2.

It is easy to see that the signaling game described above has four Nash equilibria: $\langle \sigma_1, \rho_1\rangle, \langle \sigma_2, \rho_2\rangle, \langle \sigma_3, \rho_3\rangle$, and $\langle \sigma_4, \rho_1\rangle$. But what we are interested in here are the cases where communication takes place, meaning that in different states,

different messages are sent. It is easy to see that this is the case only in the equilibria $\langle \sigma_2, \rho_2 \rangle$ and $\langle \sigma_3, \rho_3 \rangle$. In cheap talk games, the messages are not directly payoff relevant: The utility functions do not mention the messages being used. Thus, the only effect that a message can have in these games is through its information content: by changing the receiver's belief about the situation the sender (and receiver) is in. If a message can change the receiver's beliefs about the actual situation, it might also change the receiver's optimal action and thus indirectly affect both players' payoffs.

In an important article, Crawford and Sobel (1982) show that cheap talk can have real strategic impact in that it might change the receiver's optimal action but also that the amount of possible communication in cheap talk games depends on how far the preferences of the participants are aligned. They show that when the preferences are more aligned, more communication can occur through costless signaling. To put it more negatively, they show that in Lewisean cheap talk games communication is possible only if the preferences of speaker and hearer are aligned. In a zero-sum two-person game, for instance, it is predicted that communication with cheap messages is impossible: Whatever is said by the sender will be ignored by the receiver. One might think of this result as a motivation of Grice's cooperative principle, which assumes that the participants are cooperative—thus have aligned preferences—in a conversation (Grice 1967).

To establish the fact proved by Crawford and Sobel, no mention was made of any externally given meaning associated with the messages. What happens if we assume that these messages *do* have an externally given meaning, taken to be sets of situations? Thus, what happens when we adopt an externally given interpretation function "$[[\cdot]]$" that assigns to every $m \in M$ a subset of T? The interesting question is now not whether the game has equilibria in which we can associate meanings with the messages, but rather whether there exist equilibria where the messages are sent in a *credible* way. That is, are there equilibria where a speaker sends a message with meaning $\{t_i\}$ if and only if she is in state t_i? As it turns out, the old question concerning informative equilibria in signaling games without preexisting meaning and the new one concerning credible equilibria in signaling games with messages that have a preexisting meaning are closely related. Farrell (1988, 1993), Rabin (1990), Matthews et al. (1991), and Stalnaker (2006) show that costless messages with a preexisting meaning can be used to credibly transmit information only if it is known by the receiver that it is in the sender's interest to speak the truth. Communication is predicted to be possible only if the preferences are aligned. But this immediately gives rise to a *problem*. It seems that agents—human or animal—also send messages to each other, even if the preferences are less harmonically aligned. Why would they do that? In particular, how could it be that natural language is used for communication even in these unfavorable circumstances?

Reputation effects of lying in repeated games have been proposed (e.g., Axelrod and Hamilton 1981) to explain reliable communication. But experiments show that communication takes place even in one-shot games. To account for these cases, it is standardly assumed both in *economics* (starting with Spence 1973) and in *biology* (Zahavi 1975; Grafen 1990; Hurd 1995) that reliable communication is possible, if we assume that signals can be too *costly* to fake. The utility function of the sender

Table 3

	u_s, u_r	t_1	t_2	t_3	t_4
Utility	t_1	1, 0	2, −1	3, −4	4, −9
	t_2	1, −1	2, 0	3, −1	4, −4
	t_3	1, −4	2, −1	3, 0	4, −1
	t_4	1, −9	2, −4	3, −1	4, 0

takes no longer only the benefit of the receiver's action for a particular type of sender into account but also the cost of sending the message. But assuming that messages of natural languages can be costly seems counterintuitive.[1] Until now, we have not assumed that speakers are required to speak truly. Perhaps by adding this constraint, we can explain communication in more general settings. This issue is discussed in persuasion games, to which we will turn now.

2.1 Persuasion Games

Persuasion games are very similar to signaling games, but where the messages do have preexisting meaning, and it is assumed that signallers can only send true messages.

In general, we can think of a persuasion game as a game between an interested party (the sender) and a decision maker (the receiver). Let T be a finite set of *states of the world* and P a full support probability on T. The decision maker is interested in predicting the value of a *payoff relevant state* or $t_i \in T$ by choosing a state $t_j \in T$ as close as possible to the actual state t_i. The interested party's utility function u_S is strictly increasing in T. Thus, for all $t_i \in T$, $u_s(t_i, t_j) > u_s(t_i, t_k)$ just in case $j > k$. This, of course, is common knowledge, which means that the decision maker knows the ordinal preferences of the interested party. As usual, the decision maker doesn't know the actual state, but the interested party tries to persuade the decision maker that the true state is high by revealing some information. A sender strategy σ is a function from states to messages, such that for any $t \in T : t \in [\![\sigma(t)]\!]$. Thus, the set of available messages for each type, $\Omega(t)$, is a subset of $\{m \in M: t \in [m]\}$. What is important is that when the actual state is t, the sender has available a report m that rules out lower quality types. In symbols, $\forall t \in T : \exists m \in \Omega(t) : \forall t' < t : m \notin \Omega(t')$. This assumption would be satisfied, for example, if the sender could always prove the precise quality of its products or if it can prove a tight lower bound on the quality of its product. A decision maker's utility function consistent with the above assumptions can be given by $u_r(t_i, t_j) = -(j - i)^2$. This gives rise to the following type of payoff table, where the rows represent the actual states, while the columns represent the choice of the decision maker (Table 3).

We will assume that a receiver strategy is a function from messages to a probabilistic function over T, such that $\forall m \in M : \sum_t \rho(m)(t) = 1$. The identity of

[1] Though see de Jaegher (2003) for more discussion.

$\rho(m)$ will depend on what the decision maker believes, represented by probability function μ. This function μ specifies what the receiver, or decision maker, believes when the sender makes a report. Let us call a pair $\langle \rho, \mu \rangle$ a "posture" for the decision maker. Every posture requires that the decision maker forms beliefs consistent with his information and maximizes accordingly. A *naively credulous* posture is one in which the decision maker takes the sender's report at face value and simply puts $\mu(t|m) = \frac{P(t)}{P([\![m]\!])}$ for $t \in [\![m]\!]$. A *skeptical* posture $\langle \bar{\rho}, \bar{\mu} \rangle$ is one such that, for every report m, $\bar{\mu}(t_j|m) = 1$ for the *minimal* t as far as the sender is concerned, that is, $t \in [\![m]\!]$ and $\forall t' \in [\![m]\!] : u_s(\cdot, t') \geqslant u_s(\cdot, t_i)$. A skeptical posture minimizes (over all postures) the state he is going to guess. In terms of seller and buyer, a skeptical posture minimizes the quantity the buyer will purchase. Equilibria of this game are defined in terms of triples like $\langle \sigma, \rho, \mu \rangle$, where σ is a sender strategy and $\langle \rho, \mu \rangle$ is a receiver posture. The triple $\langle \sigma, \rho, \mu \rangle$ is a *sequential equilibrium* if (i) σ is the sender's best response to ρ for whatever type he is; (ii) for all m, $\rho(m)$ is the best guess of the receiver given his beliefs, and (iii) $\mu(t|m) = \frac{P(t)}{P(\sigma^{-1}(m))}$ for $t \in \sigma^{-1}(m)$ and is zero otherwise.

Milgrom and Roberts demonstrate that in such persuasion games, it is best for the decision maker to "assume the worst" about what the seller reports and that they have omitted information that would be useful (Milgrom and Roberts 1986). Their optimal equilibrium strategy will always be the *skeptical posture*. What is more, sellers will know that this is the decision maker's optimal strategy. Given this, sellers could as well reveal all they know.[2] In terms of our topic, this means that sellers/informed speakers might try to manipulate the beliefs of the decision maker by being less precise than they could be; this won't help because the decision maker will see through this attempt of manipulation. So, again, the conclusion is that standard economic theory predicts that manipulation by communication is impossible, a result that is very much in conflict with what we perceive daily.[3]

Glazer and Rubinstein have recently studied a somewhat different type of persuasion games. For them, a persuasion problem is a quadruple $\langle \{\mathrm{S}, \mathrm{H}\}, \mathrm{S}, \mathrm{A}, p, \sigma \rangle$, with speaker S, hearer H, hearer's goal A, and where p is H's probability function over S. The idea is that S wants H to do a, but H only wants to do it if the actual state s_0 is in A, $s_o \in \mathrm{A} \subseteq \mathrm{S}$. As in other persuasion games, also Glazer and Rubinstein assume that S can only use true messages. A crucial role in their games is the

[2] The argument used to prove the result is normally called the *unraveling argument*. See Jager et al. (to appear) for a slightly different version.

[3] As noted by Shin (1994), the unraveling argument is extremely sensitive to any uncertainty concerning what the informed parties *actually know*. To give a very simple example, suppose that $T = \{t_1, t_2\}$, but that the decision maker is not sure whether the sender knows the true state. Then, if the sender announces that the true state is either t_1 or t_2, the decision maker *cannot* appeal to the unraveling argument to conclude that t_1 is the true state. There is now a positive probability that the seller is genuinely uninformed and is in fact telling the whole truth. Still, one can prove a generalization of the result of Milgrom and Roberts that there always exists a sequential equilibrium $\langle \sigma, \rho, \mu \rangle$ of the persuasion game in which the disclosure strategy σ is perfectly revealing in the sense that the sender will say exactly what he knows.

persuasion function f. It is a function from messages to a number in [0, 1], where this number measures the probability that H is persuaded to do *a*. Assuming that both players are rational, S wants to choose *m* that maximizes *f*, while H wants to minimize the error probability: $\mu_{w_0}(f) = 1 - max_{m \in \sigma(s_0)} f(s)$, if $s_0 \in$ A, and $max_{m \in \sigma(s_0)} f(s)$ otherwise.

For illustration, look at the following coin toss example. This is a game about the result of five coin tosses. It is easy to see that this gives rise to 32 possible outcomes. S knows the actual outcome, but H does not. It is common knowledge that S wants H to perform *a* whatever the outcome is, but H wants to do *a* only if she is persuaded that the coin landed heads at least three times. Unfortunately, S can only inform H about the outcomes of two coin tosses. What is the optimal way for S and H to proceed? Well suppose that H's rule is to do *a* iff S demonstrates that the coin came up heads two times. In that case, there are 10 of the 32 possible outcomes where H will make the wrong choice: Do *a* although the coin came up heads only two times. Thus, the error probability is 10/32. But H can do better: He can perform *a* only if S demonstrates that the coin came up heads at two *consecutive* tosses. In this case, the error possibility is only 5/32. Glazer and Rubinstein prove that this is also the *optimal* strategy for H to choose. From our point of view we are interested in something else: Can this persuasion game perhaps explain how we try and can manipulate others? But the straightforward answer is again "no." The unique best strategy used by the hearer will always be a *skeptical* one: Always assume the worst. For instance, if the speaker would have said that the coin came up heads on the 1*st* toss and the 3*rd*, the hearer will conclude that the coin didn't come up heads in the 2*nd* trial. Manipulation can't succeed.

3 Giving Up Some Standard Assumptions

We communicate more than standard game theory predicts. This strongly suggests that standard game theory is based on some unrealistic assumptions. In this section, we will discuss three of such assumptions and indicate what might result if we give these up. First, we will discuss the assumption that what game is being played is common knowledge. Second, we will see the implications of giving up the unrealistic hypothesis that everybody is completely rational and this is common knowledge. Finally, we will discuss the assumption that our assessment of probabilities and our decisions is independent of the way the alternatives and decision problems are stated. Giving up either of these assumptions will make more room for communication and will thus be more realistic.

3.1 No Common Knowledge of the Game Being Played

In standard game theory, it is assumed that players model the game in the same way: It is common knowledge what game is played. But this seems like a highly idealized assumption. Is it not the case that players might model the game

Table 4

		b_1	b_2	b_3
Game 1. Actual game	a_1	0,2	3,3	0,2
	a_2	2,2	2,1	2,1
	a_3	1,0	4,0	0,1

Table 5

		b_1	b_2	b_3
Game 2. A thinks that B thinks	a_1	0,2	3,3	0,2
	a_2	2,2	2,1	2,1

differently or at least view others as modeling it as differently? In recent work, Feinberg (2008) demonstrates that if this possibility is taken into account, a new rationale for communication shows up. Instead of giving his theory, we will just motivate his approach by discussing one of his examples. Consider the following strategic game between row player Alice and column player Bob (Table 4).

This game has obviously exactly one Nash equilibrium: $\langle a_2, b_1 \rangle$. Standard game theory predicts that this equilibrium will be played, if it is assumed that it is common knowledge between the players that the above is indeed the game that is being played. Suppose, however, that Alice believes that Bob thinks that the only actions between which Alice can choose are a_1 and a_2, that is, Alice believes that Bob is unaware of action a_3 and thinks that the following game will be played:

In fact, however, Bob believes that it is the actual game in Fig. 1 that is being played, although he recognizes that Alice is unaware that he is considering a_3 (Table 5).

Bob also thinks that Alice thinks that Bob is considering game 2 as the actual game. Notice that although game 1 has $\langle a_2, b_1 \rangle$ as its unique Nash equilibrium, game 2 has $\langle a_2, b_2 \rangle$ as its unique equilibrium. As a result, Alice thinks it is likely that Bob will play action b_2. Alice's actual best response (i.e., in game 1) to b_2, however, is not a_2, but a_3. But because of Bob's knowledge (he is aware that Alice thinks that Bob is unaware of action a_3), he can figure out that Alice would play a_3, and his best response to this in game 1—the actual game and the game that he thinks he is playing—is action b_3. Thus, this reasoning of Alice and Bob would result in play $\langle a_3, b_3 \rangle$, which is strictly worse for both agents than the Nash equilibria play in either game.

Suppose that before the agents make their choice, agents are allowed to send a message. We have seen that in standard game theory, pre-play communication can normally be ignored (the messages are not credible) if the preferences of the agents are not well aligned. On the other hand, if it were common knowledge that the game was game 1, for example, pre-play communication would be ignored as well because that game has only one Nash equilibrium. We will see that in our case, however, pre-play communication makes perfect sense. Bob can send a message ("I know you can also play a_3") which makes clear to Alice that he is aware of action a_3. It is immediately clear that this message is credible: There is no reason for Alice to think she is being manipulated. As a result, it becomes common knowledge

that it is actually game 1 that is being played and that Alice should thus choose a_1 instead of a_3. Together with Bob's best response, we end up with the Nash equilibrium $\langle a_2, b_1 \rangle$ which gives rise to a higher utility for Alice and for Bob. Thus, it was indeed rational for Bob to communicate as he did. Feinberg (2008) discusses more cases like this.

Thus, we can explain more cases of rational communication than we could before if we don't make the ideal, but unrealistic assumption that it is always common knowledge which game is being played.

3.2 No Common Knowledge of Rationality

A Nash equilibrium is the solution concept in game theory, but it is not always easy to reach it. In quite a number of games, however, a simple procedure will do: (iterated) elimination of strategies that violate the canons of rationality, that is, that are strongly dominated. In case we end up with exactly one (rationalizable) strategy for each player, this strategy combination must be a Nash equilibrium. This procedure crucially depends, however, on a very strong epistemic assumption: *common knowledge of rationality*; not only must every agent be ideally rational, everybody must also know of each other that they are rational, and they must know that they know it and so on ad infinitum. It is harder to justify Nash equilibria in general, but also such a justification leans heavily on this strong assumption. Unfortunately, there exists a large body of evidence that the assumption of common knowledge of rationality is highly unrealistic.

The p-beauty contest game (Moulin 1986) is based on a similar game by Keynes (1936) and was introduced to highlight how unrealistic this assumption is. In this game, each of $n > 2$ players chooses a whole number between 0 and 100. Let us say that k is the average of these n numbers. The winners of the game are those players who choose their numbers closest to $2k/3$, and they share the prize equally. Obviously, you shouldn't choose any number greater than $\frac{2}{3} \times 100 \approx 67$, because such a strategy has payoff 0, whereas the mixed strategy playing 0–67 with equal probability has a strictly positive payoff. Thus, any of the former strategies is strongly dominated by the latter mixed strategy and should thus be eliminated after one round of eliminating strongly dominated strategies. A second round of eliminating strongly dominated strategies, however, eliminates choices above $\frac{2^2}{3} \times 100 \approx 44$ in a similar way. Continuing in this manner, we see that the only strategy that is not eliminated in any round is the strategy to choose is 0. Experimental evidence, however, shows that this would be a very poor choice. Working with various groups of size 14–16, Nagel (1995) found that the average number chosen was 35, which is between two and three rounds of iterated elimination of strongly dominated strategies. Thus, in this game, we cannot assume common knowledge of rationality: Agents "think ahead" only a very limited number of rounds.

In the previous sections, we have seen that deception and manipulation could not be explained within standard game theory. One reason for this is that it assumes common knowledge of rationality. If it is common knowledge that everybody is rational, any attempt of deception will be anticipated, and the anticipation thereof will be anticipated as well and so on *ad infinitum*. But we have seen above that it is not in accordance with experimental evidence to assume common knowledge of rationality. Is it possible to explain deception and manipulation if we give up this assumption?

Indeed, it can be argued that wherever we do see attempted deceit in real life, we are sure to find at least a belief of the deceiver (whether justified or not) that the agent to be deceived has some sort of limited reasoning power that makes the deception at least conceivably successful. Some agents are more sophisticated than others and think further ahead. To model this, one can distinguish different *strategic types* of players. A strategic type captures the level of strategic sophistication of a player and corresponds to the number of steps that the agent will compute in a sequence of iterated best responses. One can start with unstrategic level-0 players. An unstrategic level-0 hearer (a credulous hearer), for example, takes the semantic content of the message he receives literally, and doesn't think about why a speaker used this message. Obviously, such a level-0 receiver can sometimes be manipulated by a level-1 sender, as we have seen in Section 1. But such a sender can in turn be "seen through" by a level-2 receiver if she understood why the level-1 sender sent what he sent, etc. In general, a level-k + 1 player is one who plays a best response to the behavior of a level-k player. (A *best response* is a rationally best reaction to a given belief about the behavior of all other players.) A fully sophisticated agent is a level-inf player. In a very interesting article, Crawford (2003) shows that in case sender and/or receiver believes that there is a possibility that the other player is less sophisticated than he is himself, deception is possible. Moreover, even sophisticated players can be deceived if they are not sure that the opponent is fully rational or not. Crawford assumed that messages have a specific semantic content, but did not presuppose that speakers can only say something that is true. It is possible, however, to use the same kind of idea to show that manipulation is possible in such circumstances as well in persuasion games as discussed in Section 2.1 of this chapter.

We can conclude that (i) it is unnatural to assume common knowledge of rationality, and (ii) by giving up this assumption, we can explain much better why people communicate than standard game theory can: Sometimes we communicate to manipulate others on the assumption that the others don't "see it through," that is, that we are smarter than them (whether this is justified or not).

3.3 Framing and Reference-Point-Based Preferences

Although our standard of living increased a lot the last decades, psychological research on happiness finds that subjective measures of well-being are relatively stable over time. This suggests that one's well-being crucially depends on the value

of one's own properties compared to that of others. In more abstract terms, *utility* is *reference-based*, which is in contrast with the additive utility function underlying standard game theory. The most natural reference point to compare one's welfare is the current *status quo* position. Now, psychologists have discovered that people value payoffs according to whether they are *gains* or *losses* compared to their current status quo position. Subjective well-being is associated not so much with the *level* of income, but more with *changes* of income. Moreover, agents are much more averse to lose X euros than that they are attracted to winning X euros. These phenomena can be illustrated by the following famous Asian disease experiment due to Tversky and Kahneman (1981).

In the two versions of this experiment, which takes the form of a questionnaire, a separate but similar population was confronted with the following hypothetical scenario: "Imagine that the USA is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed."

In version 1 of the experiment, subjects were offered the choice between programs A and B, which are described as follows: "If program A is adopted, 200 people will be saved. If program B is adopted, there is 1/3 probability that 600 people will be saved and 2/3 probability that no people will be saved."

In version 2 of the experiment, subjects were offered the choice between programs C and D: If program C is adopted, 400 people will die. If program D is adopted, there is 1/3 probability that nobody will die and 2/3 probability that 600 people will die.

When the choice is between A and B, 72% of the subjects choose A; when the choice is between C and D, 78% choose D. This is in spite of the fact that, from the perspective of expected utility maximization, the two examples are perfectly equivalent. Apparently, the experimenter, by framing the example in a different manner, can influence the reference point of the subject and can cause a preference reversal.

Let the US population have size X before the outbreak, and let us assume that the decision maker is an expected utility maximizer with an increasing, strictly concave Bernouilli utility function $u(\cdot)$ over the post-outbreak US population. Then it is easy to see that the decision maker should not make any difference between versions 1 and 2. The expected utility of programs A and C is equally $u(X - 400)$. The expected utility of programs B and D is $\frac{1}{3} \times u(X) + \frac{2}{3} \times u(X - 600)$. Note that the numbers are chosen such that if for any Y we have $u(Y) = U$, then $u(X - 400) = \frac{1}{3} \times u(X) + \frac{2}{3} u(X - 600)$. It follows that as soon as $u(\cdot)$ is strictly concave, then $u(X - 400) > \frac{1}{3} \times u(X) + \frac{2}{3} \times u(X\,600)$ so that programs A and C should be preferred. But this is contradicted by the results of Kahneman and Tversky's experiment.

In order to account for such choices counter to expected utility theory (in this and other experiments), Kahneman and Tversky construct *prospect theory*. The elements of this theory are that decision makers think in terms of gains and losses with respect to an exogenously given reference point. Decision makers are risk

averse with respect to gains and risk loving with respect to losses (reflection effect). They are loss averse, in that, for example, they are hurt more by a 100 loss than they enjoy a 100 gain. Finally, they overweigh small probabilities.

We only need the reference point and the reflection effect to account for the choices in the Asian disease problem. Let r be the reference point of the decision maker, in this case, a reference post-outbreak US population. Consider a strictly concave valuation function $v(\cdot)$, which is defined both over gains and losses, and with $v(0) = 0$. In the loss region, for a post-outbreak population of Y, the consumer's utility then takes the form $v(Y - r)$ if $Y \geq r$ (gains region) and takes the form $-v(Y - r)$ if $Y < r$ (loss region). The consumer's utility is then indeed strictly convex in the loss region and strictly concave in the gains region. For the rest, all is the same as in expected utility theory.

Assume now that in version 1 of the experiment, the reference point is that nobody is saved so that any person saved is seen as a gain. It follows that $r = X - 600$ and that we are everywhere in the gains region. In this case, the decision maker prefers program A if and only if:

$$\upsilon(X - 400 - (X - 600)) > \frac{1}{3}\upsilon(X - (X - 600)) + \frac{2}{3}\upsilon((X - 600) - (X - 600)) \text{ if}$$

$$\upsilon(200) > \frac{1}{3} \times \upsilon(600) \text{ if and only if}$$

$$3 \times \upsilon(200) > \upsilon(600).$$

It is clear that this is valid as soon as we have a strictly concave $v(\cdot)$ so that the decision maker's utility is strictly concave in the gain region.

Assume that in version 2 of the experiment, the reference population is that nobody dies so that any person who dies is seen as a loss. It follows that $r = X$ and that we are everywhere in the loss region. The decision maker now prefers program C if and only if

$$-V(X - (X - 400)) < -\frac{1}{3}V(X - X) - \frac{2}{3}V(X - (X - 600)) \text{ if and only if}$$

$$V(400) > \frac{1}{3}V(600) \text{ if and only if}$$

$$V(600) < 1.5V(400).$$

Again, this is valid as soon as we have a strictly concave $v(\cdot)$ so that the decision maker's utility is strictly convex in the loss region.

But why would the reference points be different in version 1 and version 2? It seems that a reference point is induced merely by expressing the news as a gain ("are saved") or as a loss ("die") with respect to a reference population. In version 1, by expressing the news as a gain with respect to a reference population where 600 people are killed, the decision maker is induced to be risk averse. In version 2, by expressing the news as a loss with respect to a reference population the decision maker is induced to be risk averse.

Ducrot (1973) and Anscombre and Ducrot (1983) have argued that we have to look at language use from an *argumentative perspective* to be able to explain the appropriate use of certain adversarial connectives. Merin (1999) sought to provide a formal analysis of their insights, but failed (cf. van Rooij 2004). We believe that a more appropriate formalization is possible making use of prospect theory. The idea is that the argumentative function of an adversary connective used by a manipulative persuader is to suggest a reference point with respect to which the main body of information given should be compared. Consider the following modified statements for the two versions. In version 1*, the first sentence is now stated as a lack of a gain. Further, the adversarial connective "still" induces a contrast with this situation of no gain. By their nature, such adversarial connectives would seem to invite the listener to make comparisons and so to think in terms of gains and losses. For the rest, all the populations are stated as gains. In version 2*, the first sentence is clearly stated as a loss. The adversarial connective "however" contrasts this with situations where the losses are smaller. All further populations are expressed as losses. Further, the order in which programs C and D are expressed is reversed in comparison to the original experiment. The order in which the alternatives are expressed may also induce a reference point.

Thus, an empirical question here lies in the extent to which adversarial connectives and expressions suggesting gains and/or losses, and the order in which alternatives are presented, are successful in creating reference points with listeners.

Version 1*: Imagine that the USA is preparing for the outbreak of an unusual Asian disease. If no program is adopted, there will be no rescue for 600 people. Still, if program A is adopted, 200 people will be saved, and if program B is adopted, there is 1/3 probability that all 600 people will be saved and 2/3 probability that none of them is saved.

Version 2*: Imagine that the USA is preparing for the outbreak of an unusual Asian disease. If we fail to interfere, 600 people will lose their lives. However, if program B is adopted, there is 1/3 probability that none of these people will die and a 2/3 probability that all 600 of them will continue to die. If program A is adopted, 400 of these people will die.

It should be noted that applications of prospect theory are not confined to uncertainty. Another example (due to Anscombre and Ducrot 1983), which does not involve uncertainty and where both the order of the statements and adversarial connectives seem to play a role, is the following. Consider a restaurant critic who objectively observes a restaurant to be both more expensive than other good restaurants and better than other expensive restaurants.

Version 1^{**} The restaurant is expensive, but good.
Version 2^{**} The restaurant is good, but expensive.

Each time, the earliest statement may induce the reference point. Version 1^{**} could induce a reference point with the decision maker of considering restaurants as expensive. Yet, among expensive restaurants, the restaurant is one of the good ones. Version 2^{**} could induce as a reference point that restaurants serve good food. Yet, among restaurants serving good food, the restaurant at hand is expensive. If the

decision maker reads the critics review of the restaurant, and considers eating home as a choice with a utility of zero, then in version 1^{**}, the decision maker would decide to go to the restaurant (as she perceives positive utility in going to the restaurant), and in version 2^{**}, she would prefer to stay at home (as she perceives a negative utility in going to the restaurant).

4 Conclusion

So, why do we talk so much? Perhaps because our preferences are much aligned and participants of a conversation all profit from a larger distribution of knowledge. This would be the ideal picture, but we doubt it is the true reason behind (all) our talking. We also talk if our preferences are not aligned. No, we talk so much, we argue, because, among others, (i) we think we know better in which situation we are than others (3.1), (ii) we think we are smarter than others (3.2), or (iii) we think we can influence the probabilities and utilities of others by the way we frame their decision problems. In short, we talk and argue so much because we believe others are *bounded rational agents*.

References

Anscombre, J. C., & Ducrot, O. (1983). *L'Argumentation dans la Langue*. Brussels: Mardaga.

Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science, 411*, 1390–1396.

Crawford, V. (2003). Lying for strategic advantage: rational and boundedly rational misrepresentations of intentions. *American Economic Review, 93*, 133–149.

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

de Jaegher, K. (2003). Error-proneness as a handicap signal. *Journal of Theoretical Biology, 224*, 139–152.

Ducrot, O. (1973). *La preuve et le dire*. Paris: Mame.

Farrell, J. (1988). Communication, coordination and Nash equilibrium. *Economic Letters, 27*, 209–214.

Farrell, J. (1993). Meaning and credibility in cheap-talk games. *Games and Economic Behavior, 5*, 514–531.

Feinberg, Y. (2008). Meaningful talk. In K. Apt & R. van Rooij (Eds.), *New perspectives on games and interaction* (pp. 105–120). Amsterdam: Amsterdam University Press.

Franke, M. (2009). *Signal to act. Game theory in pragmatics*. PhD thesis, University of Amsterdam.

Glazer, J., & Rubinstein, A. (2008). A study in the pragmatics of persuasion: A game theoretical approach. In K. Apt & R. van Rooij (Eds.), *New perspectives on games and interaction* (pp. 121–140). Amsterdam: Amsterdam University Press.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology, 144*, 517–546.

Grice, H. P. (1967). *Logic and conversation*. Typescript from the William James Lectures, Harvard University (Published in Grice, P. (1989), *Studies in the way of words*. Cambridge, MA: Harvard University Press, pp. 22–40).

Horn, L. (1989). *A natural history of negation*. Chicago: University of Chicago Press.

Hurd, P. (1995). Communication in discrete action-response games. *Journal of Theoretical Biology, 174*, 217–222.
Keynes, J. M. (1936). *The general theory of employment, interest and money*. New York: Harcourt Brace and Co.
Lewis, D. (1969). *Convention*. Cambridge, MA: Harvard University Press.
Matthews, S., Okuno-Fujiwara, M., & Postlewaite, A. (1991). Refining cheaptalk equilibra. *Journal of Economic Theory, 55*, 247–273.
Merin, A. (1999), *Die Relevanz der Relevanz: Fallstudie zur formalen Semantik der englishen Konjunktion 'but'* (Arbeitspapiere SFB 340, nr. 142), Stuttgart: Stuttgart University.
Milgrom, P., & Roberts, J. (1986). Relying on the information of interested parties. *Journal of Economics, 17*, 18–32.
Moulin, H. (1986). *Game theory for the social sciences* (2nd ed.). New York: NYU Press.
Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review, 85*, 1313–1326.
Rabin, M. (1990). Communication between rational agents. *Journal of Economic-Theory, 51*, 144–170.
Spence, M. (1973). Job market signalling. *Quarterly Journal of Economics, 87*, 355–374.
Stalnaker, R. (2006). Saying and meaning, cheap talk and credibility. In A. Benz, G. Jäger, & R. van Rooij (Eds.), *Game theory and pragmatics*. Basingstoke: Palgrave MacMillan.
Tversky, A., & Kahneman, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choices. *Science, 211*, 453–458.
van Rooij, R. (2004). Cooperative versus argumentative communication. *Philosophia Scientia, 2*, 195–205.
Zahavi, A. (1975). Mate selection – A selection for a handicap. *Journal of Theoretical Biology, 53*, 205–214.