Tjalling C. Koopmans Research Institute

Discussion Paper Series nr: 11-25

Game-theoretic pragmatics under conflicting and common interests

Kris De Jaegher
Robert van Rooij

Universiteit Utrecht

Utrecht School of Economics

Tjalling C. Koopmans Research Institute
Utrecht School of Economics
Utrecht University

Kriekenpitplein 21-22
3584 TC Utrecht
The Netherlands
telephone +31 30 253 9800
fax +31 30 253 7373
website www.koopmansinstitute.uu.nl

The Tjalling C. Koopmans Institute is the research institute and research school of Utrecht School of Economics.
It was founded in 2003, and named after Professor Tjalling C. Koopmans, Dutch-born Nobel Prize laureate in economics of 1975.

In the discussion papers series the Koopmans Institute publishes results of ongoing research for early dissemination of research results, and to enhance discussion with colleagues.

Please send any comments and suggestions on the Koopmans institute, or this series to J.M.vanDort@uu.nl

ontwerp voorblad: WRIK Utrecht

How to reach the authors

Please direct all correspondence to the first author.

Kris De Jaegher
Utrecht University
Utrecht School of Economics
Kriekenpitplein 21-22
3584 TC Utrecht
The Netherlands.
E-mail: k.dejaegher@uu.nl

Robert van Rooij
Universiteit van Amsterdam
Institute of Logic, Language and Computation (ILLC)
Oude Turfmakrt 143
1012 GC Amsterdam
The Netherlands
E-mail: r.a.m.vanrooij@uva.nl

Utrecht School of Economics
Tjalling C. Koopmans Research Institute
Discussion Paper Series 11-25

Game-theoretic pragmatics under conflicting and common interests*

Kris De Jaegher[a]
Robert van Rooij[b]

[a]Utrecht University School of Economics
Utrecht University

[b]Institute of Logic, Language and Computation
University of Amsterdam

December 2011

Abstract
This paper combines a literature overview of existing literature in game-theoretic pragmatics, with new models that fill some voids in the literature. We start with an overview of signaling games with a conflict of interest between sender and receiver, and show that the literature on such games can be classified into models with direct, costly, noisy and imprecise signals. We then argue that this same subdivision can be used to classify signaling games with common interests, where we fill some voids in the literature. For each of the signaling games treated, we show how equilibrium-refinement arguments and evolutionary arguments can be interpreted in the light of pragmatic inference.

Keywords: Signaling games, pragmatics, equilibrium refinements, evolutionary game theory.

JEL classification: D82, D83.

Acknowledgements
We wish to thank participants of the workshop Game Theory and Communication: Prospects and Syntheses, May 28 and 29, 2010, for helpful comments.

1. Introduction

It is safe to say that the use of game theory has become more common within the field of pragmatics. While such game-theoretic pragmatics focuses specifically on signaling games, the field lacks an overview of the signaling games to which pragmatics can be applied. This is not a surprise, as such overviews of signaling games are lacking in game theory itself. The purpose of this article is two-fold. The *first* purpose is to provide an overview of the different signaling games to which pragmatics can be applied. As game theory and in particular economics has mostly been interested in conflicts of interest, we start by giving an overview of the different types of signaling games involving a *conflict of interest* treated in the literature. We identify direct signaling games, costly signaling games, noisy signaling games and imprecise signaling games. Each time, the purpose of these games is to show how the use of signals of a particular form can solve the conflict of interest between sender and receiver, in ensuring that the receiver finds the sender's signals credible in spite of the conflict of interest between them.

Yet, given Grice's (1967) cooperative principle, pragmatics is more interested in signaling games with *common interests*. As we show, the four categories of direct signaling games, costly signaling games, noisy signaling games and imprecise signaling games are also very worthy of an analysis under common interests. This brings us to our *second* purpose, namely of filling some gaps in the literature. In particular, direct signaling games with common interests and noisy signaling games with common interests have to our knowledge not been treated in game-theoretic pragmatics, and we attempt to fill this gap.

The common theme in the paper is that all of the signaling games have multiple Nash equilibria, and the question is whether we can find arguments in the literature on refinements of the Nash equilibrium so that only the efficient Nash equilibrium is selected. The most obvious way in which it is clear that there are multiple Nash equilibria is that all of the games we treat have both pooling equilibria, where the sender does not send any signals (or sends signals, but not in an informative manner) and the receiver does not obtain any information, and separating equilibria, where the sender does sender does send signals and where the receiver does receive information. The question arising then is: if the receiver expects that the pooling equilibrium will be played, but unexpectedly receives a certain signal, will the receiver make some inference that the signal could only have been sent by a particular player? This is the question arising when applying equilibrium refinements to signaling games. As this is also a question of pragmatic inference, game-theoretic pragmatics thus seems to coincide with such equilibrium refinement arguments. This is familiar for costly signaling games, but we will show that it also applies for direct, noisy, and imprecise signaling. We argue that these games can inspire the formulation of pragmatic rules that seem sensible.

Less obviously, some of the games we treat have both efficient separating equilibria and inefficient separating equilibria. The question arises again whether equilibrium refinement arguments allow us to select against these inefficient separating equilibria. In general, the answer is positive for conflict-of-interest signaling games, but no for common-interest games. In this sense, in predicting efficient play, common-interest games are more problematic than conflict-of-interest games. This is because the typical argument in conflict-of-interest signaling games is that signals of a particular form can solve the conflict of interest. For instance, if a sender is biased to always report the same state of the world, this can be solved by sending a signal that is prohibitively costly to send in one of the states. Inefficiency can then arise because the same purpose can be achieved with a cheaper signal. But if a receiver can realize that a signal with a particular but high cost level is credible, then he will also

realize this of a signal that is a bit cheaper. So with a conflict of interest, the elimination of inefficient separating equilibria is not problematic.

In common-interest games, however, typically in an inefficient separating equilibrium, there is a complete reversal of meaning compared to the efficient separating equilibrium. For instance, whereas in an efficient separating equilibrium a costly signal refers to an infrequent state, in an inefficient separating equilibrium it refers to a frequent state. Once sender and receiver play an inefficient separating equilibrium, we do not have arguments why they would move away from this equilibrium, exactly because this would require a complete reversal of meaning. We argue, however, that one should take as one's starting point the pooling equilibrium. After all, if we want to understand how people learn to communicate, we must take as a starting point a situation where they do not communicate. The problem is that, unlike what is the case in conflict of interest games, equilibrium refinements in some cases predict that the *inefficient* separating equilibrium will be played.

In Section 2, we present a general signaling model, of which all models treated in this paper are variants. Here, we also introduce the several equilibrium refinements that we treat in this paper. In Section 3, we treat direct, costly, noisy and imprecise signaling models for the case of a conflict of interest between sender and receiver. In Section 4, we treat the four corresponding signaling models when there are common interests between sender and receiver. We end with a discussion in Section 5.

2. Workhorse model and equilibrium refinements

The game has two players, namely the sender Z, and the receiver H. At stage 1 of the game, Nature decides which state from a set of states $\mathbb{T} = \{L, M, R\}$ occurs, with typical element labeled T, where Nature's choice is observed by Z but not by H. State T occurs with probability π_T. If two states of the world are sufficient for our purposes, we will assume that $\pi_R = 0$. For brevity, we sometimes refer to a sender who observes state T as a sender of type T.

At stage 2, for each state $T = L, M, R$, Z decides to send a signal from the set $\mathbb{S}_T$ with typical element labeled S_T. Any such set always includes the possibility for the sender of doing nothing, so of not sending any signals. The set of signals that the sender is able to send may differ between one state of the world to the other, in which case we have a *revelation* model of signaling, as signals may then directly reveal aspects of the state of the world, as certain signals can only be sent in certain states of the world. In this case, we speak of *direct signals*, in that signals directly reveal information about the state of the world. For instance, a fruit seller that pretends to sell excellent apples may let consumers taste an apple, thus revealing the quality of the apple. In most of our analysis, however, we will assume that the set $\mathbb{S}$ is the same for each state of the world (every signal can be sent in every state of the world), in which case the subscript is dropped. We then have *indirect signals*, as the signals do not directly contain or reveal information about the state of the world (independent of the quality of her apples, the fruit seller can pretend that they are very good). For indirect signals, we both consider the case where the cardinality $|\mathbb{S}|$ is equal to or larger than the number of states of nature that occur with positive probability minus one, so that a separating equilibrium is possible, and the case where $|\mathbb{S}|$ is smaller than the number of states of nature minus one (i.e. one signal is only available though there are three states), in which case a fully separating equilibrium is not possible. In the latter case, we have *imprecise signaling*, as the signaling system can never be precise about all states of the world.

Sending a signal S after having observed state T comes at cost $C(S|T)$ to Z, which may differ according to T. We talk of *costless signaling* if $\forall S = L, M, R$, $\forall T \in \mathbb{T}$: $C(S|T) = 0$. We talk of *costly signaling* when there is at least one signal S and state T such that $C(S|T) > 0$.

Indirect signals may or may not have a commonly understood meaning. We say that players have *no common language* if there is no commonly-known meaning of the signals. This means that meaning can only arise in equilibrium. We say that players have a *common language* if signals have a commonly-known meaning. This means that the set $\mathbb{S}$ consists of a set of propositions, with commonly-known meanings such as "state T occurs", or "either state T_1 occurs or state T_2", etc. When assuming that there is a common language, we immediately assume that this common language is rich, in that one is able to state every possible proposition about the states of the world in it. It should be stressed that the existence of a common language does not mean that the Z uses signals truthfully, or that the receiver H trusts these signals to be used truthfully. The fact that signals have common meaning therefore does not mean that signals are credible.

At stage 3, Nature decides whether or not H receives the signal sent by Z. In particular, H receives signal S with probability $\mu(S|S)$, and does not receive the signal with probability $\mu(0|S)$, where $\mu(0|S) \geq 0$, $\mu(S|S) + \mu(0|S) = 1$. We talk of *noisy signaling* when there is at least one signal S such that $\mu(0|S) > 0$, and of *noiseless signaling* when $\forall S: \mu(0|S) = 0$. We will consider noisy signaling only for the case of indirect signals.

At stage 4, H decides which action, with typical action labeled A, to take from a set of actions $\mathbb{A}$, where $\mathbb{A} = L, M, R$. To make the model as parsimonious as possible, in part of our analysis we assume that doing R is a strictly dominated action to H, so that effectively only actions L and M are relevant.

At stage 5, for $I = Z, H$; $A = L, M, R$; $T = L, M, R$, player I receives payoff $U_I(A|T)$ (where for Z, the costs of sending signals should still be subtracted). It is the case that $U_H(A_1|T) > U_H(A_2|T)$ for $A_1 = T$, $A_1 \neq A_2$, so that the receiver prefers that the action with the same label as the state of nature is taken, rather than an action with a different label. If it is alo the case that $U_Z(A_1|T) > U_Z(A_2|T)$ for $A_1 = T$, $A_1 \neq A_2$, we have a game with *common interests*. If not, we have a game with a *conflict of interest*.

A Nash equilibrium of the signaling game is a pair consisting of a signaling strategy $S^*(T)$ and of an action strategy $A^*(S)$ such that these strategies are mutual best responses. As is typically the case for signaling games, our game has multiple Nash equilibria. This can already be seen by the fact that it has pooling equilibria. Simply, if H decides to take the same action A whether or not a signal is received, then it is best response for Z not to send any signals; this in turn makes the given strategy of H a best response.

Given the multiplicity of Nash equilibria, we now define a set of equilibrium refinements that have been defined specifically in the context of signaling games. We start with two refinements that are rooted in a rational approach, where each time H is assumed to reason about Z's intentions when observing Z deviate from a given equilibrium. We first define Farrell's (1993) *neologism proofness*, which is only relevant for indirect signals (where the signal does not directly reveal anything about the state of the world, as every signal can be sent in every state of the world) that have a commonly-known meaning. Whenever we refer to neologism-proofness in what follows, this means that we assume a commonly-known language. While the issue of the *meaning* of signals is then resolved, this does not necessarily mean that signals are used honestly, i.e. that they are *credible*. Intuitively, if H unexpectedly

observes an out-of-equilibrium signal, i.e. a "neologism", if this neologism has a commonly known meaning, and if only certain types of senders have an incentive to send it, then *H* will respond to it in a certain manner, which may then induce certain sender types to indeed send such a neologism. If this is so, then the considered equilibrium is not neologism-proof.

Definition 1 (Farrell, 1993). Consider an equilibrium *E**, and consider an out-of-equilibrium signal S'' taking the form of a proposition *P* from a commonly-known language. Let A_x be *H*'s best response when *P* is truthful, and let *Z* in *E** send signal S' (which includes the possibility of not sending any signal) in state T'. Denote by $A^*(S)$ the receiver's response to signal *S* in equilibrium *E**. Then we say that *E** is *not neologism-proof* iff

$$\forall T \notin P:$$
$$\left[\mu(S|S)U_Z(A^*(S)|T)+\mu(0|S)U_Z(A^*(0)|T)-C(S|T)\right] >$$
$$\left[\mu(S''|S'')U_Z(A_x|T)+\mu(0|S'')U_Z(A^*(0)|T)-C(S''|T)\right]. \quad (1)$$

$$\forall T' \in P:$$
$$\left[\mu(S'|S')U_Z(A^*(S')|T')+\mu(0|S')U_Z(A^*(0)|T')-C(S'|T')\right] \leq$$
$$\left[\mu(S''|S'')U_Z(A_x|T')+\mu(0|S'')U_Z(A^*(0)|T')-C(S''|T')\right].$$[1] ■ (2)

Our next two equilibrium refinements are found in Cho and Kreps (1987), and are only relevant for cheap and costly signaling. Whenever we refer to these equilibrium refinements, the assumption is that there is no common language. Whereas neologism proofness only deals with the credibility of signals, these refinements both deal with meaning and credibility of messages. Specifically for costly signaling, it may be the case that for some *T* that can be observed by *Z*, *Z* never wants to send a signal *S*, whatever *H*'s response to it (including whatever response *H* may have when no signal is observed). It is reasonable then to assume that *H* should never interpret *S* as having come from *T*.

Definition 2 (Cho and Kreps, 1987). Let

$$\min_{A_v, A_w}\left[\mu(S|S)U_Z(A_v|T)+\mu(0|S)U_Z(A_w|T)-C(S|T)\right] >$$
$$\max_{A_x, A_y}\left[\mu(S'|S')U_Z(A_x|T)+\mu(0|S')U_Z(A_y|T)-C(S'|T)\right]. \quad (3)$$

Then we say that for type *T*, signal S' is *strictly dominated* by signal *S*. ■

From equation (3), it is immediately clear that the strict domination only applies to costly signals: with costless signals, the worst payoff that can be obtained by sending signal *S* can never be better than the best payoff obtained by sending signal S'.

The manner in which we can now apply the concept of strictly dominated strategy is by repeatedly eliminating them from the game. E.g., we may first note that certain types of senders would never want to send a particular signal. We can then first eliminate these signals from the senders signal sets, so that analytically we obtain a game that resembles a direct signaling game. In this restricted game, we can then further eliminate receiver strategies that do not take into account that only certain types of senders can have send certain signals.

[1] Note that the response when no signal arrives is always the same as in the equilibrium, as the receiver can then not observe that the sender has deviated from the equilibrium.

Rather than checking whether there are certain sender types who would never want to send certain signals in any circumstances, we can also look at whether it is not the case that certain types would never have an incentive to deviate from a particular equilibrium. This is the idea of the *intuitive criterion* (Cho and Kreps, 1987). The idea is similar to the one of neologism-proofness, except that it is not required now that out-of-equilibrium signals have commonly-known meaning, thus the intuitive criterion deals both with meaning and credibility. Suppose that players expect to play Nash equilibrium *E**. It may now be the case that for some *T* that can be observed by *Z*, whatever the action taken by *H*, *Z* always does worse by sending signal S' than in *E**. Additionally, assume that types other than *T* exist that do have an incentive to deviate from *E**. Then *H* should interpret such a signal S' as having come from such types. The Nash equilibrium does not meet the intuitive criterion then. We add to Cho and Kreps' original treatment the definition of a separating equilibrium that *destabilizes* a pooling equilibrium. We then select out separating equilibria that do not destabilize any pooling equilibrium. The argument then is that, if the purpose is to explain how communication evolves, we must take the pooling equilibrium as a starting point (see Van Rooij (2008) for this argument). A separating equilibrium that cannot be achieved from previous or at least expected play of the pooling equilibrium is then eliminated.

Definition 3.

(i) (Cho and Kreps, 1987) Consider a Nash equilibrium *E** that includes equilibrium, strategies $S^*(T)$, $A^*(S)$. Define *BR*(.) as the set of best responses to a set of types. Define as *T*(*S*) the set of types in which signal *S* can be sent. Consider the set τ of all *T* such that

$$\left[\mu(S|S)U_Z(A^*(S)|T)+\mu(0|S)U_Z(A^*(0)|T)-C(S|T)\right] > \max_{A_x\in BR[T(S'')]}\left[\mu(S''|S'')U_Z(A_x|T)+\mu(0|S'')U_Z(A^*(0)|T)-C(S''|T)\right], \quad (4)$$

We then say that signal S'' is *equilibrium dominated* for type *T*. Then a Nash equilibrium does not meet the *intuitive criterion* if there exists a type T' sending a signal S' in *E**, such that

$$\left[\mu(S'|S')U_Z(A^*(S')|T')+\mu(0|S')U_Z(A^*(0)|T')-C(S'|T')\right] < \min_{A_y\in BR(\neg\tau)}\left[\mu(S''|S'')U_Z(A_y|T')+\mu(0|S'')U_Z(A^*(0)|T')-C(S''|T')\right]. \quad (5)$$

(ii) We say of equilibrium *E*** that it *destabilizes* equilibrium *E** if a type-signal pair (S',T') as described under (i) exists for *E**, and the strategy consisting of *Z*'s best response in *E**, with the type-signal pair (S',T') changed into (S'',T'), is identical to his best response in *E***.

∎

It is clear from equation (4) that the intuitive criterion can be applied only if signals are direct, or if signals are indirect and costly. With direct signals, the cost of revealing is assumed zero, but as pointed out under the maximization sign, it may simply be impossible for certain signals to have come from certain senders, which the receiver takes into account in his best response. With indirect, costly signals, the same effect may be caused by the cost of sending certain signals to certain types.

One interpretation of both the Nash equilibrium, and of equilibrium refinements of the Nash equilibrium, is that players should not be interpreted as being literally as rational as they are modeled to be. The point is that by trial and error, they can learn to behave *as if* they were rational. For this reason, we also introduce an equilibrium selection criterion that is rooted in evolutionary game theory, and thus in learning. We do not explicitly model evolutionary dynamics and keep our evolutionary criterion unsophisticated, in simply look at whether through evolutionary drift (where players may switch to alternative strategies, if these strategies are equally best responses; see Binmore and Samuelson, 1999) and some evolutionary process such as replicator dynamics or best-response dynamics (where players play a strategy more often the better it did in the previous period), there is an evolutionary path from Nash equilibrium E^* to Nash equilibrium E^{**}, but not back from Nash equilibrium E^{**} to Nash equilibrium E^*. For instance, in a pooling equilibrium, it is a weak best response for receivers to start reacting to an out-of-equilibrium signal in the same manner as in a separating equilibrium. Evolutionary drift can then lead them to do this. If a sufficient number of receivers change their behavior in this manner, the senders will then next learn to behave as in the separating equilibrium. The proportion of receivers that needs to drift in a particular direction may be large. Following Sobel (1993), we then simply accept that it may take a very long time before the equilibrium is destabilized. This is contrary to e.g., recently Pawlowitsch (2008), who does not allow for such drift, and for this reason obtains that sender and receiver may get stuck in inefficient equilibria.

Definition 4. Consider a Nash equilibrium E^* with strategy profile $\left[S^*(T), A^*(S)\right]$. Consider the set $BR_H\left[S^*(T)\right]$ of best responses of H to $S^*(T)$. Consider a Nash separating equilibrium E^{**} with strategy profile $\left[S^{**}(T), A^{**}(S)\right]$. Then we say that E^{**} is *attainable* from E^* if the set of strategies $BR_H\left[S^*(T)\right]$ contains a strategy $A'(S)$ to which $S^{**}(T)$ is a best response.

We finally note an equilibrium refinement criterion that is hard to formalize, namely that certain equilibria may be "focal" (Schelling (1960)). A particular equilibrium may be conspicuous to players, it may be common knowledge that it is conspicuous to them, and for this reason they may each expect that all other players will play according to it. In the context of signaling games, it may be focal to use a certain signal in a particular manner.

3. Conflict of interest

In all the models that we treat in this section, Z has a bias towards action L, and thus as such wants to pretend that state L occurs even if this is not true. How can Z now still credibly signal that it is optimal for R to take action L? In all the models below, Z who observes L makes his signal credible by making the sending of this signal prohibitively costly to any other type of Z. In the direct-signaling model (3.1), signals directly reveal information, and for a sender who did not observe state L it is impossible to reveal information that suggests that state L occurs – whereas a sender who does observe state L has no reason to hide information. In the costly-signaling model (3.2), the signal literally takes the form of incurring a cost that other types would never want to incur. In the noisy-signaling model (3.3), sending signals as such may be costless, but Z makes signal L credible by making it noisy in such a way that it becomes unattractive to send for other types. In the imprecise model (3.3), again signals may be costless, but Z makes signal L less precise, by sending this signal both in states L and M. In

this manner, it becomes too expensive for a Z who observed R to cheat, as the induced action is then too remote from Z's optimal action. In each of the models, H infers the intentions of Z by noting that the signal sent could only have been sent by a Z who observed L . While the underlying principle is thus each time the same, the principle is put to different uses in each of the models.

3.1 Direct signaling

In a first type of signaling that we treat, we can abstract both from the problem of the meaning of signals, and from their credibility. With *direct signals*, sending a signal means that information that unambiguously determines the state of the world can directly be revealed, or that a cue that makes it more likely that a certain state of the world occurs can be revealed. The underlying mechanisms can be explained here using only two states M and L. For instance, let the sender be a dictator who may (state M) or may not (state L) have weapons of mass destruction, and is always better off if the receiver concludes that she does not have such weapons (action L). If the dictator can simply reveal whether or not she has weapons, the receiver may infer that if she does not reveal anything, she must have weapons (action M). This is because the dictator who does not have weapons might as well reveal this, since this information is to her advantage (Milgrom, 1981). The principle of pragmatic inference applied here is that a sender who does not reveal her information must have something to hide.

In a more subtle version of this model, related to the model of Glazer and Rubinstein (2001), because of the costs of providing evidence, the dictator can only either reveal whether or not she has nuclear weapons, and whether or not she has chemical weapons, but not both. An additional problem here is that the dictator now has an incentive to only reveal information that is to her advantage. Thus, when she has nuclear weapons but not chemical weapons, she reveals that she does not have chemical weapons, but remains quiet about nuclear weapons. Realizing this, the receiver may not find a revelation that one of the types of weapons is missing credible. A way around this is that sender and receiver coordinate on an equilibrium where information is revealed about only one of the types of weapons. E.g., the receiver concludes that the dictator does not have weapons of mass destruction if she reveals not to have nuclear weapons, but concludes that she does have weapons if she reveals not to have chemical weapons. This is not because revealing that one does not have a certain type of weapons is objectively more convincing, but because it stops the dictator from only revealing information that is to her advantage. The principle of pragmatic inference applied here is that a sender is credible if she reveals a random piece of information, rather than having picked one that is to her advantage. We first treat the model where all information can directly be revealed, and then the model where only part of the information can be revealed.

Direct signaling conflict-of-interest model with one cue underlying the states of the world. In the simplest version of this model, there are only two states of the world M and L, and the sender in state M (L) can either reveal a *single* cue that unambiguously shows that state M (L) occurs, or reveal nothing. When nothing is revealed, the receiver of course cannot distinguish whether this was done in state M or state L. As shown in Proposition 1, in terms of the weapons example, a pooling equilibrium exists where nothing is revealed. Depending on the parameters, the receiver may then either interpret that there are (action M) or are no (action L) weapons of mass destruction. At the same time, a separating equilibrium exists where the dictator reveals when she has no weapons (state L), and does not reveal otherwise (state M). This is a best response for the dictator if the receiver interprets failure to reveal

information as evidence of possession of mass destruction (action *M*). The dictator who does have weapons weakly prefers not to reveal her information.

Proposition 1. Consider the direct-signaling game with two states of the world, where the sender can without costs reveal each of the states of the world. Then a Pareto-efficient separating equilibrium exists where the sender reveals state *L* when state *L* occurs, and where state *M* is not revealed. If $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} \leq \frac{\pi_L}{\pi_M}$, the pooling equilibrium is also Pareto-efficient. If $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} > \frac{\pi_L}{\pi_M}$, the pooling equilibrium is not Pareto efficient.

Proof:

Given that the receiver responds *L* when *L* is revealed, and as revelation is costless, the sender in state *L* may as well reveal *L*. If the sender in state *L* reveals the state of the world, by not revealing the sender in state *M* is immediately identified by the receiver.

If $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} \leq \frac{\pi_L}{\pi_M}$, in the pooling equilibrium action *L* is always taken, which is the preferred outcome to each type of sender. As the separating equilibrium makes the sender in state *M* worse off, this pooling equilibrium is Pareto-efficient.

If $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} > \frac{\pi_L}{\pi_M}$, in the pooling equilibrium action *M* is always taken, so that each type of sender becomes weakly better off by moving to the separating equilibrium.

QED

As a pooling equilibrium always exists, we now check whether equilibrium selection arguments select the separating equilibrium.

Proposition 2. Consider the direct-signaling game with one cue underlying the states of the world, where the sender can without costs reveal each of the states of the world.

Let $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} > \frac{\pi_L}{\pi_M}$. Then:

(i) *Pooling equilibria* (*Pareto inefficient*): do not survive iterated elimination of strictly dominated strategies and do not meet intuitive criterion, do not destabilize any separating equilibria and cannot be attained from them.

(ii) *Separating equilibria (Pareto efficient)*: survive iterated elimination of strictly dominated strategies, meet intuitive criterion, destabilize pooling equilibria and can be attained from them.

Let $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} < \frac{\pi_L}{\pi_M}$. Then:

(iii) *Pooling equilibria* (*Pareto efficient*): survive iterated elimination of strictly dominated strategies, meet intuitive criterion, but do not destabilize any separating equilibria and cannot be attained from them.

(iv) *Separating equilibria (Pareto efficient)*: survive iterated elimination of strictly dominated strategies, meet intuitive criterion; do not destabilize pooling equilibria but can be attained from them

Proof:

(i) It is strictly dominated for *H* to interpret a revealed state *L* as *M*. By iterated elimination of strictly dominated strategies, given this fact, it is a best response for *Z* to reveal state *L*. Further, the pooling equilibrium does not meet the intuitive criterion because it is impossible for type *M* to reveal a state *L*, so that any revealed state *L* can only come from type *L* (equation (5)). It follows that the separating equilibrium destabilizes the pooling equilibrium. Finally, it is a best response in the pooling equilibrium for the receiver to do *L* upon out-of-equilibrium revelation of state *L*. The best response to this strategy by the receiving is the strategy of the separating equilibrium, which is therefore attainable from the pooling equilibrium.

(ii) It is strictly dominated for the receiver to respond to a revealed state *M* with action *L*. Yet, in the separating equilibrium where only type *L* reveals, type *M* does not have any incentive to reveal her state. It follows that this separating equilibrium survives iterated elimination of strictly dominated strategies. For the same reason, this separating equilibrium meets the intuitive criterion, so that it is not destabilized by any other equilibria. The same applies to separating equilibria where both states are revealed.

(iii) Even though it is strictly dominated for revealed states not to be met with the right response, given the form of the pooling equilibrium neither type has an incentive to reveal their types. Applying equation (4), consider S'' as revealing one's type. Then type *M* is better off by not revealing, so that set τ contains only state *M*. However, equation (5) cannot be applied because type *L* only has a weak incentive to reveal. The pooling equilibrium therefore meets the intuitive criterion and is not destabilized by any separating equilibrium. Finally, in the pooling equilibrium it is a best response for the receiver to respond in the appropriate way to revealed signals. Given the resulting receiver strategy, it is a weak best response for type *L* to reveal, and a strict best response for type *M* not to reveal. But this is the equilibrium strategy in a separating equilibrium, which is therefore attainable from the pooling equilibrium.

(iv) The only difference with (ii) is that separating equilibria do not destabilize pooling equilibria, which was shown under (iii).

QED

Intuitively, if in the pooling equilibrium the receiver concludes that the dictator has weapons of mass destruction, then the pooling equilibrium is destabilized by iterated elimination of strictly dominated strategies and by the intuitive criterion because the receiver knows that the revealed fact that the dictator does not have weapons must be truthful, and because the dictator who does not have weapons has an incentive to reveal this fact. However, if in the pooling equilibrium the receiver concludes that the dictator does *not* have weapons of mass destruction, then the dictator does not have any incentive to reveal that she does possess them, so that the pooling equilibrium is not destabilized.

Direct signaling conflict-of-interest model with two cues underlying the states of the world. We now consider an extension where it continues to be the case that there are only two states of the world *M* (has weapons) and *L* (does not have weapons), but where there are *two* cues that decide whether state *M* or state *L* occurs. This extension bears resemblance to the mechanism described by Glazer and Rubinstein (2001), in that only one of the two cues will turn out to be convincing in equilibrium. Both the first cue (does or does not have nuclear weapons) and the second cue (does or does not have chemical weapons) take on either a value of 0 ("no weapons") or 1 ("weapons"). Each cue is determined independently, where for each cue a value of 0 occurs with probability $(1-x)$. It follows that event (0, 0), occurs with probability $(1-x)^2$, events (1, 0) and (0, 1) with probability $x(1-x)$, and event (1, 1) with probability x^2. In event (0, 0), response *L* is optimal (interpret the dictator as not having

weapons), so that we can say that state *L* occurs with probability $\pi_L = (1-x)^2$. In events (1, 1), (1, 0), (0, 1) the best response is *M* (interpret the dictator as having weapons), so that in this event, state *M* occurs, where $\pi_M = 2x(1-x) + x^2$.

If the sender is able to reveal both cues, then we have a similar model as before. The receiver then interprets any non-revealed cue as having a value of 1. The reasoning is that a receiver who observes a 0 may as well reveal it. Here, we assume instead that the sender is only able to reveal a *single* cue. Given that the sender prefers that action *L* is always taken, a strategy of the sender could be to always reveal a cue 0 if at least one cue 0 occurs, so that the sender always reveals information that is to her advantage. Under the assumptions of Proposition 3, however, it is a best response then for the receiver to always do *M* (interpret as having weapons) in this case. Under these same assumptions, a separating equilibrium still exists where the sender only reveals one of the cues, say the first one, when it has a value of one, and never reveals the second cue.

Intuitively, if the sender is only able to reveal one cue at a time, then in order for the sender to make it credible that she is not only revealing information on the cue that is to her advantage, she should reveal only a *particular* cue. In terms of the dictator example, e.g., the dictator should only reveal whether or not she has nuclear weapons.

Proposition 3. In the direct-signaling model with two cues underlying the state of the world, let $\frac{(1-x)}{2x} < \frac{\left[U_H(M|M) - U_H(L|M)\right]}{\left[U_H(L|L) - U_H(M|L)\right]} < \frac{(1-x)}{x}$. Then in the Pareto-inefficient pooling equilibrium, *M* is always done. In the Pareto-efficient separating equilibrium, the sender only reveals a particular one of the cues if it has value zero, where the other cue is always maintained unrevealed.

Proof:

Let the sender reveal a zero whenever he observes at least one zero. Then the receiver prefers to do *M* given that

$$\frac{2x(1-x)U_H(M|M) + (1-x)^2 U_H(M|L)}{2x(1-x) + (1-x)^2} > \frac{2x(1-x)U_H(L|M) + (1-x)^2 U_H(L|L)}{2x(1-x) + (1-x)^2}$$

$$\Leftrightarrow$$

$$\frac{\left[U_H(M|M) - U_H(L|M)\right]}{\left[U_H(L|L) - U_H(M|L)\right]} > \frac{(1-x)}{2x}$$

It follows that there is no separating equilibrium where the sender reveals a 0 when it is to her advantage.

The sender can instead employ a strategy where she only reveals whether the value of a particular cue out of the two cues is 0, and never reveals anything about the other cue. If the receiver observes a 0 revealed for the particular cue, it is a best response for her to do *L* iff

$$\frac{x(1-x)U_H(L|M) + (1-x)^2 U_H(L|L)}{x(1-x) + (1-x)^2} > \frac{x(1-x)U_H(M|M) + (1-x)^2 U_H(M|L)}{x(1-x) + (1-x)^2}$$

$$\Leftrightarrow$$

$$\frac{(1-x)}{x} > \frac{\left[U_H(M|M) - U_H(L|M)\right]}{\left[U_H(L|L) - U_H(M|L)\right]}$$

When nothing is revealed, it is a best response for the receiver to do *M*, as given the sender's equilibrium strategy, this automatically means that the particular cue takes on a value of 1. Further, given that the sender does not reveal anything about the other cue, it is a weak best response for the receiver to do *L* whenever the other cue is revealed (independently of whether a 0 or a 1 is revealed).

In turn, given that the receiver follows such a strategy, it is a weak best response for the sender not to reveal anything about the other cue, and to reveal nothing when the particular cue has value 1. However, given that the sender has a preference for response *L* being taken, she reveals when the particular cue has a value of zero.

In the pooling equilibrium, response *M* is adopted as $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} > \frac{\pi_L}{\pi_M}$, where $\frac{\pi_L}{\pi_M} = \frac{(1-x)^2}{2x(1-x)+x^2}$. This is because we have assumed that $\frac{(1-x)}{2x} < \frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]}$, and because $\frac{(1-x)}{2x} > \frac{(1-x)^2}{2x(1-x)+x^2}$. The pooling equilibrium is Pareto-inefficient as each player is better off in the specified separating equilibrium.
QED

Proposition 4 again checks whether pooling equilibria are eliminated by equilibrium selection arguments.

Proposition 4. Consider the direct-signaling game with two cues underlying the states of the world, let $\frac{x}{2(1-x)} < \frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} < \frac{x}{(1-x)}$.

(i) The *pooling equilibrium* survives iterated elimination of strictly dominated strategies, meets the intuitive criterion, does not destabilize the separating equilibrium, and cannot be attained from the separating equilibrium.

(ii) The *separating equilibrium* survives iterated elimination of strictly dominated strategies, and meets the intuitive criterion. It does not destabilize the pooling equilibrium, but can be attained from it.

Proof:

Consider a pooling equilibrium, and let the receiver unexpectedly observe a revealed "0". Then it is not the case that only a sender who observed (0, 0) could benefit from deviating from the pooling equilibrium. Thus, the pooling equilibrium meets the intuitive criterion, and the separating equilibrium does not destabilize it.

The separating equilibrium meets the intuitive criterion for the following reason. Suppose that in equilibrium only the first cue is revealed, but the receiver still unexpectedly observes a revealed "0" from the second cue. Then it is not the case that only a sender who observed (0, 0) could benefit from deviating from the separating equilibrium. Yet, the separating equilibrium can be attained from the pooling equilibrium because the receiver could become predisposed to considering only one of the two possible revealed 0's as conclusive.
QED

In the model with two cues (possession of nuclear weapons, and of chemical weapons), in terms of the dictator example, in the pooling equilibrium the receiver concludes that the dictator has weapons of mass destruction. Even though only a single cue can be revealed, a

separating equilibrium still exists if only evidence on the non-possession of a *single* type of weapons is considered relevant. Yet, the problem is that if in the pooling equilibrium the receiver unexpectedly observes the sender to reveal that she does not have a certain type of weapons, there is no way of telling that she still did not pick only the information that is to her advantage, and that she still possesses weapons of the other type. Thus, from this perspective the mechanism that only one piece of evidence is considered relevant does not seem to work. Still, if in the pooling equilibrium through drift the receiver becomes predisposed to consider only evidence on e.g. nuclear weapons as conclusive, then the separating equilibrium can still evolve from the pooling equilibrium.

The principle of considering only one piece of evidence relevant also works if one piece of evidence is somehow considered as focal. Rubinstein and Glazer (2001) offer the following motivating example:

"You are participating in a public debate about the level of education in the world's capitals. You are trying to convince the audience that in most capital cities, the level of education has risen recently. Someone is challenging you, bringing up indisputable evidence showing that the level of education in *Bangkok* has deteriorated. Now it is your turn to respond. You have similar, indisputable evidence to show that the level of education in *Mexico City***,** *Manila***,** *Cairo* and *Brussels* has gone up. However, because of time constraints, you can argue and present evidence only about one of the four cities mentioned above. Which city would you choose for making the strongest counterargument against *Bangkok*?"

The authors conducted an experiment on this, showing that most of the time the counterargument is Manila. The point here is that by the fact that Manila is closest to Bangkok, the sender signals that she is choosing a particular cue, and not necessarily the cue that is most to her advantage. Thus, there is some focal feature of one of the cues that makes it seem that a particular cue was chosen, and not one necessarily to the advantage of the sender.[2] Such an argument cannot be caught by iterated elimination of weakly dominated strategies, or by the intuitive criterion.

3.2 Costly signaling

In the next model, signals do not directly reveal the sender's information, so that credibility is an issue. Meaning may or may not be an issue, in that signals may or may not be grounded in a common language. The sender always would like to get the same response. The manner in which signals are still made credible is that signals are costly, and that the cost incurred from sending a signal differs according to one's type. The principle of pragmatic inference applied here is that a particular signal could never have been sent by certain sender types, as sending such a signal is too costly to them. For instance, suppose that a sender always wants to be considered as being rich. Then lighting a cigar with a $100 bill is a credible signal that one is rich, as a poor person finds it too costly to do this. Lighting a $100 bill may or may not already have a commonly known meaning of signaling richness. The point here is that this signal is credible. This model has independently been developed in economics (educational signaling on the labor market, Spence, 1973) and in biology (handicap signaling, Zahavi, 1974).

[2] It is easy to generalize the model we present to make it closer to Rubinstein and Glazer. In such a model, there are more than two cues determine the state of the world, and the sender can only communicate two cues. The sender can could convince the receiver that she is not only reporting two cues that are to her advantage if these two cues have focal features, such as lying close to each other.

Costly signaling conflict-of-interest model. In the simplest version of this model, there are only two states *M* and *L*. A set of signals is available to the sender, which are differentiated according to their costs. However, a costly signal can still credibly signal that state *L* occurs.

Proposition 5. Consider the conflict-of-interest costly signaling game. Then a separating Nash equilibrium where the sender sends signal *S* in state *L*, and no signal in state *M*, exists iff $C(S|M) \geq U_Z(L|M) - U_Z(M|M)$ and $\frac{C(S|M)}{U_Z(L|M) - U_Z(M|M)} \geq \frac{C(S|L)}{U_Z(L|L) - U_Z(M|L)}$. The separating equilibrium with the signal *S* from set $\mathbb{S}$ which has the lowest $C(S|L)$ such that the conditions are valid, is the only separating equilibrium that is Pareto-efficient. If $\frac{[U_H(M|M) - U_H(L|M)]}{[U_H(L|L) - U_H(M|L)]} \leq \frac{\pi_L}{\pi_M}$, the pooling equilibrium is also Pareto-efficient. If $\frac{[U_H(M|M) - U_H(L|M)]}{[U_H(L|L) - U_H(M|L)]} > \frac{\pi_L}{\pi_M}$, the pooling equilibrium is not Pareto efficient.

Proof:
If *Z* plays the strategy of the specified candidate separating equilibrium, it is a best response for *H* to do *L* when receiving signal *S*, and to do *M* otherwise. If *H* plays the strategy of the specified candidate separating equilibrium, it is a best response for *Z* to send signal *S* in state *L* and not to send any signal in state *M* iff $U_Z(M|M) \geq U_Z(L|M) - C(S|M)$ and $U_Z(L|L) - C(S|L) \geq U_Z(M|L)$. The conditions follow. QED

The condition $C(S|M) \geq U_Z(L|M) - U_Z(M|M)$ means that it must be costly for *Z* to send signal *S*, meaning that costly signaling is a necessary condition for a separating equilibrium in the specified game. It should be noted that the signal need not be costly to type *L*, meaning that in equilibrium a cost of signaling need not ever be incurred (Hurd, 1995). In the most well-known version of the second condition for the existence of a separating equilibrium, $[U_Z(L|M) - U_Z(M|M)] = [U_Z(L|L) - U_Z(M|L)]$, so that the condition reduces to $C(S|M) > C(S|L)$. This means that the signal must be *differentially costly*; put otherwise, it must be more costly to send signal *S* in state *M* than in state *L* (Grafen, 1990). Yet, less well-known is that signals need not be differentially costly. If $C(S|M) = C(S|L)$, so that signal *S* is equally costly to all types, then a separating equilibrium can still exist if $[U_Z(L|M) - U_Z(M|M)] > [U_Z(L|L) - U_Z(M|L)]$. This is the *differential benefits* version of Spence-Zahavi model (Johnstone, 1997). Intuitively, either there is a signal that is too costly to send in state *M*; or signal cost is the same in all states, but the sender in state *L* is more motivated to send a costly signal.

In the set of separating equilibria, clearly the Pareto efficient equilibrium is the one where the signaling cost incurred by the sender is as low as possible. But this need not be the only Pareto-efficient equilibrium. This is because, as pointed out by Spence, a separating equilibrium does not always make *Z* better off. In particular, if the game has a pooling equilibrium where *L* is always done, *Z* is worse off in the separating equilibrium, where his preferred action *L* does not always get done, and where additionally a costly signal may have to be sent. It follows that both the efficient separating equilibrium and the pooling equilibrium are Pareto efficient in the case. If in the pooling equilibrium *M* is always done, then only efficient separating equilibrium is Pareto-efficient. We next check which of the multiple equilibria survive equilibrium refinements.

Proposition 6. Consider the conflict-of-interest costly signaling game.

Let $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} > \frac{\pi_L}{\pi_M}$.

(i) The *pooling equilibria* (*Pareto inefficient*) are not neologism proof against separating equilibria, do not survive iterated elimination of strictly dominated strategies, do not meet the intuitive criterion, do not destabilize any separating equilibria and cannot be attained from them.

(ii) The *inefficient separating equilibria* are not neologism proof against efficient separating equilibria, do not survive iterated elimination of strictly dominated strategies, do not meet the intuitive criterion, destabilize pooling equilibria but not efficient separating equilibria, can be attained from pooling equilibria but not from efficient separating equilibria.

(iii) The *efficient separating equilibria* is neologism proof, survives iterated elimination of strictly dominated strategies, meets the intuitive criterion, destabilizes inefficient separating equilibria and pooling equilibria, and can be attained from them.

Let $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} < \frac{\pi_L}{\pi_M}$. Then:

(iv) The *pooling equilibria* (*Pareto efficient*) are neologism proof, survive iterated elimination of strictly dominated strategies, meet the intuitive criterion, but do not destabilize any separating equilibria and cannot be attained from them.

(v) The *inefficient separating equilibria* are not neologism proof against efficient equilibria, do not survive iterated elimination of strictly dominated strategies, do not meet intuitive criterion, do not destabilize efficient separating equilibria or pooling equilibria, and cannot be attained from them.

(vi) The *efficient separating equilibria* is neologism proof, survives iterated elimination of strictly dominated strategies, meets the intuitive criterion, destabilizes inefficient separating equilibria but not pooling equilibria, and can be attained from inefficient separating equilibria but not from pooling equilibria.

Proof:

(i) If the costly signal already has a common meaning "do *L*", if the receiver follows it, type *M* does not want to send it by $U_Z(M|M) \geq U_Z(L|M) - C(S|M)$ (equation (1)), but *L* does by $U_Z(L|L) - C(S|L) \geq U_Z(M|L)$ (equation (2)), so that the pooling equilibrium is not neologism proof. By $U_Z(M|M) \geq U_Z(L|M) - C(S|M)$ it is strictly dominated for *Z* to send a signal in state *M*, even if the receiver responds with *L*. The strategy where a signal is sent in state *M* can thus be eliminated, so that the receiver interprets any signal without a common meaning as referring to state *L*. Given $U_Z(M|M) \geq U_Z(L|M) - C(S|M)$ it is further equilibrium dominated for type *M* to send a signal (equation (4)). It follows by equation (5) that the pooling equilibrium does not meet the intuitive criterion, and is destabilized by the efficient separating equilibrium.

(ii) By the same reasoning as under (i), a cheaper costly signal for which the assumptions are still valid makes an inefficient separating equilibrium not neologism proof. For all such signals, it is strictly dominated to send them in state *M*, and equilibrium dominated with respect to the inefficient separating equilibrium to send them in state *M*. It follows that an inefficient separating equilibrium is eliminated by iterated elimination of strictly dominated strategies, does not meet the intuitive criterion, and is destabilized by the efficient separating equilibrium.

(iii) The efficient separating equilibrium is neologism proof and meets the intuitive criterion because the sender in state M does not have any incentive to send a signal, and because the sender in state L does not have any incentive to send a more expensive signal. Thus, the efficient separating is not destabilized by any other equilibria. Other equilibria are not attainable from it, because even if the receiver also responds to more expensive signals, the sender will not send them.
(iv) The pooling equilibrium is neologism proof because both types M and L are already receiving their preferred response at no cost, and have no incentive to send a costly signal. Even though it is strictly dominated for type M to send a signal so that by iterated elimination of strictly dominated strategies the response to a signal is always L, given the form of the pooling equilibrium type L does not have any incentive to send a signal. The pooling equilibrium meets the intuitive criterion because applying equation (4) the set τ contains both states of the world, so that equation (5) cannot be applied. The pooling equilibrium is therefore not destabilized by any separating equilibrium. Finally, even if in the pooling equilibrium it is a best response for the receiver to interpret an out-of-equilibrium signal as coming from type L, it is not a best response for the sender to send the signal, so that no separating equilibrium is attainable from the pooling equilibrium.
(v) The only difference with (ii) is that inefficient separating equilibria do not destabilize pooling equilibria and cannot be attained from them, which was shown under (iv).
(vi) The only difference with (iii) is that inefficient separating equilibria do not destabilize pooling equilibria and cannot be attained from them, which was shown under (iv).
QED

Concluding, Proposition 6 tells us the following in terms of the $100 bill example. As only a rich person has an incentive to burn $100, a receiver will always interpret such a signal to have come from a rich person. This is independent of whether or not burning $100 already refers to richness in a commonly known signaling system. For the same reason, in equilibrium the sender will not burn more money than is necessary to prove that she is rich. Nevertheless, if the pooling equilibrium is such that the receiver without information interprets the sender to be rich (e.g., because there are few poor people), then while a separating equilibrium exists, it will not be played if players initially play the pooling equilibrium. The sender is already interpreted to be rich without wasting $100, and therefore does not send the signal.

3.3 Noisy signaling

Using a story adapted from Blume, Board and Kawamura (2007) (see De Jaegher and Van Rooij, 2011), Romeo wants to know from Juliet whether or not she loves him. Romeo can either conclude that Juliet loves him, conclude that she does not love him, or make no conclusion. Whether or not she loves Romeo, Juliet most prefers that Romeo concludes that she loves him, and least prefers that he does not make any conclusion, with Romeo's conclusion that he does not love her in between. Because of Juliet's preferences, her signal "I love you", expressed in a commonly known language, is not credible to Romeo. Still, as argued, Juliet's signal "I love you" can be made credible if it is ambiguous. In particular, if her signal "I love you" is sometimes misinterpreted by Romeo to be inconclusive, and is misinterpreted in this manner more often than her "I don't love you" signal, then Juliet may still prefer to be honest. Juliet may achieve this by sending the "I love you" signal in several contexts, where Romeo imperfectly observes these contexts. She may e.g., send this signal both in the context where she loves Romeo, and in the context where she does not love him, but just had an excellent time. Romeo gets an imperfect cue of whether Juliet had an excellent

time, and makes no conclusion about her feelings if she says "I love you" and he gets a cue that she had an excellent time. In this way, Juliet's signal "I love you" sometimes leads to Juliet's least preferred response. The principle of pragmatic inference applied here by Romeo is that only a Juliet who really loves him would be willing to make her signal so ambiguous.

Conflict-of-interest noisy signaling model (De Jaegher, 2003a, 2003b)**.** In order to construct a simple model reflecting this intuition, let there be two states of the world *L* and *M*, and three responses *L*, *M* and *R* (related models where a conflict of interest is solved by means of noisy signals can be found in Myerson, 1991; Farrell, 1993; Blume, Board and Kawamura, 2007; Blume and Board, 2009). In each state of the world, the receiver prefers the right response, but prefers response *R* when getting too little information. The sender in each state prefers response *L* to response *M* to response *R*. The sender can choose from a set of costless, noisy signals which each get lost with different probabilities. Denote by $\mu(0|S)$ the probability that a signal gets lost, and by $\mu(S|S)$ the probability that it arrives. The receiver is assumed to know what level of noise is attached to a received signal *S*.

Proposition 7. Consider the conflict-of-interest noisy signaling model. Let $[U_Z(M|M)-U_Z(R|M)]*\ [U_Z(L|L)-U_Z(R|L)] \geq [U_Z(M|L)-U_Z(R|L)]\ *[U_Z(L|M)-U_Z(R|M)]$ and let $[U_H(R|L)-U_H(M|L)]*[U_H(R|M)-U_H(L|M)] \geq [U_H(L|L)-U_H(R|L)]*[U_H(M|M)-U_H(R|M)]$. Then a separating Nash equilibrium where *Z* sends signal S_1 in state *L* and signal S_2 in state *M* and where the receiver does *R* when not receiving a signal, exists only if $\mu(0|S_1)>\mu(0|S_2)>0$. A unique Pareto-efficient equilibrium exists with a minimal level of noise on both signals. Finally, in the pooling equilibrium, the receiver does *R*.

Proof:

When not receiving any signal, *H* should prefer to do *R* to *L*:

$$\frac{\pi_L\mu(0|S_1)}{\pi_L\mu(0|S_1)+\pi_M\mu(0|S_2)}U_H(R|L)+\frac{\pi_M\mu(0|S_2)}{\pi_L\mu(0|S_1)+\pi_M\mu(0|S_2)}U_H(R|M)\geq$$
$$\frac{\pi_L\mu(0|S_1)}{\pi_L\mu(0|S_1)+\pi_M\mu(0|S_2)}U_H(L|L)+\frac{\pi_M\mu(0|S_2)}{\pi_L\mu(0|S_1)+\pi_M\mu(0|S_2)}U_H(L|M)$$
$$\Leftrightarrow$$
$$\pi_M[U_H(R|M)-U_H(L|M)]\mu(0|S_2)\geq\pi_L[U_H(L|L)-U_H(R|L)]\mu(0|S_1) \quad (6)$$
$$\Leftrightarrow$$
$$\frac{\mu(0|S_2)}{\mu(0|S_1)}\geq\frac{\pi_L[U_H(L|L)-U_H(R|L)]}{\pi_M[U_H(R|M)-U_H(L|M)]}, \quad (7)$$

where the right-hand side in (7) is smaller than 1. It follows from (7) that it needs to be the case that $\mu(0|S_2)>0$. Also, when not receiving any signal, *H* should prefer to do *R* to *M*:

$$\frac{\pi_L\mu(0|S_1)}{\pi_L\mu(0|S_1)+\pi_M\mu(0|S_2)}U_H(R|L)+\frac{\pi_M\mu(0|S_2)}{\pi_L\mu(0|S_1)+\pi_M\mu(0|S_2)}U_H(R|M)\geq$$
$$\frac{\pi_L\mu(0|U)}{\pi_L\mu(0|S_1)+\pi_M\mu(0|S_2)}U_H(M|L)+\frac{\pi_M\mu(0|S_2)}{\pi_L\mu(0|S_1)+\pi_M\mu(0|S_2)}U_H(M|M)$$
$$\Leftrightarrow$$

$$\mu(0|S_1)\pi_L\left[U_H(R|L)-U_H(M|L)\right] \geq \mu(0|S_2)\pi_M\left[U_H(M|M)-U_H(R|M)\right] \quad (8)$$

$$\Leftrightarrow$$

$$\frac{\pi_L\left[U_H(R|L)-U_H(M|L)\right]}{\pi_M\left[U_H(M|M)-U_H(R|M)\right]} \geq \frac{\mu(0|S_2)}{\mu(0|S_1)}, \quad (9)$$

where the left-hand side in (9) is larger than 1. By (6) and (8), it should be the case that $\left[U_H(R|L)-U_H(M|L)\right] > 0$ and $\left[U_H(R|M)-U_H(L|M)\right] > 0$. Combined with (7) and (9), it follows that a necessary condition for the existence of a separating equilibrium is that $\left[U_H(R|L)-U_H(M|L)\right]*\left[U_H(R|M)-U_H(L|M)\right] \geq \left[U_H(L|L)-U_H(R|L)\right]*\left[U_H(M|M)-U_H(R|M)\right]$.

At the same time, the sender should prefer to send the right signal in each state, and prefer this to not sending any signal at all. That is, for *Z* who observes state *L*, we must have:

$$\mu(S_1|S_1)U_Z(L|L)+\mu(0|S_1)U_Z(R|L) \geq \mu(S_2|S_2)U_Z(M|L)+\mu(0|S_2)U_Z(R|L) \quad (10)$$

and

$$\mu(S_1|S_1)U_Z(L|L)+\mu(0|S_1)U_Z(R|L) \geq U_Z(R|L) \quad (11)$$

For *Z* who observes state *M*, we must have:

$$\mu(S_2|S_2)U_Z(M|M)+\mu(0|S_2)U_Z(R|M) \geq \mu(S_1|S_1)U_Z(L|M)+\mu(0|S_1)U_Z(R|M) \quad (12)$$

and

$$\mu(S_2|S_2)U_Z(M|M)+\mu(0|S_2)U_Z(R|M) \geq U_Z(R|M) \quad (13)$$

By the fact that $U_Z(M|L) > U_Z(R|L)$, $U_Z(L|M) > U_Z(R|M)$, constraints (11) and (13) are slack. Constraints (10) and (12) can respectively be rewritten as:

$$\left[1-\mu(0|S_1)\right]\left[1-\mu(0|S_2)\right]^{-1} \geq \left[U_Z(M|L)-U_Z(R|L)\right]\left[U_Z(L|L)-U_Z(R|L)\right]^{-1} \quad (14)$$

and

$$\left[U_Z(M|M)-U_Z(R|M)\right]\left[U_Z(L|M)-U_Z(R|M)\right]^{-1} \geq \left[1-\mu(0|S_1)\right]\left[1-\mu(0|S_2)\right]^{-1} \quad (15)$$

In (14), the right-hand side is smaller than 1; in (15), the left-hand side is also smaller than 1. From the latter, it follows that in any separating equilibrium, it must be the case that $\mu(0|S_1) > \mu(0|S_2)$. Further, it follows from (14) and (15) that a separating equilibrium can only exist if $\left[U_Z(M|M)-U_Z(R|M)\right]*\left[U_Z(L|L)-U_Z(R|L)\right] \geq \left[U_Z(M|L)-U_Z(R|L)\right]*\left[U_Z(L|M)-U_Z(R|M)\right]$. It follows that under the conditions specified in the proposition, levels of $\mu(0|S_1)$ and $\mu(0|S_2)$ exist such that constraints (7), (9), (14) and (15) are valid, so that a separating equilibrium exists.

The fact that there is a unique Pareto-efficient separating equilibrium with minimal noise can be seen by plotting (7), (9), (14) and (15) in graph with $\mu(0|S_1)$ on one axis and $\mu(0|S_2)$ on the other. The fact that R is done in the pooling equilibrium follows from (7) or (9) combined with the result that $\mu(0|S_1) > \mu(0|S_2)$.

QED

We next look at whether equilibrium refinement arguments select a separating equilibrium, and in particular a Pareto-efficient one. We do not consider iterated elimination of strictly dominated strategies. As there are no direct costs of sending the signals, it is not true that certain types would never want to send them. By the same reasoning, it is not true that only certain types can get better than in a given equilibrium be sending them, so that the intuitive criterion is also not considered here.

Proposition 8. Consider the conflict-of-interest noisy signaling model.

(i) The *pooling equilibria* are neologism proof and cannot be attained from any separating equilibria.

(ii) The *inefficient separating equilibria* are not neologism proof against efficient separating equilibria, and can be attained from pooling equilibria but not from efficient separating equilibria.

(iii) The *efficient separating equilibrium* is neologism proof and can be attained from both inefficient separating equilibria and pooling equilibria.

Proof:

(i) A receiver who expects the pooling equilibrium to be played and unexpectedly observes a noisy signal S_1 or a noisy signal S_2 cannot make any inference that these signals must have been sent by a certain type of sender. Only if both signals are used at the same time may they be credible, and the receiver has no indication that the sender is using both signals at the same time. Still, in a pooling equilibrium, it is a best response for the receiver to respond by L (M) to a noisy S_1 (S_2) signal, so that the sender's strategy from a separating equilibrium becomes a best response. It follows that any separating equilibria are attainable from the pooling equilibrium.
We next look at the attainability of a pooling equilibrium from a separating equilibrium. While in a separating equilibrium R is already played when not receiving any signal, it is not a best response for the sender to deviate and stop sending signals. It follows that pooling equilibria are not attainable from separating equilibria.

(ii) By plotting (7), (9), (14) and (15) in graph with $\mu(0|S_1)$ on one axis and $\mu(0|S_2)$ on the other, it can be seen that starting from a Pareto-inefficient separating equilibrium, a Pareto superior separating equilibrium exists where one of the signals is made less noisy. Thus, if starting from an inefficient separating equilibrium, a slightly less noisy signal with a commonly known meaning is observed, then by (1) it cannot have been used untruthfully. Given that it is less noisy, by (2) it will be sent.
In a Pareto-efficient separating equilibrium, it is a best response to the receiver to also find more noisy out-of-equilibrium signals credible, but this does not make it a best response for the sender to send them instead. Thus, Pareto-inefficient separating equilibria are not attainable from efficient ones.

(iii) In the efficient separating equilibrium, while by (1) only a truthful sender may have an incentive to send a more noisy signal so that a more noisy signal is trusted, this does not give the sender any incentive to send a more noisy signal. Further, in the Pareto-inefficient equilibrium, it is a best response for the receiver to find the signals from the efficient separating equilibrium credible, and this makes it a best response for the sender to send them. Thus, the Pareto-efficient separating equilibrium is attainable from inefficient separating equilibria.

QED

In terms of the Romeo-Juliet example, the point of Proposition 8 is that a single noisy signal does not make Juliet's signals credible. Even if her "I love you" signal is sometimes misinterpreted, if the starting point is the pooling equilibrium, not only a Juliet who loves, but one who does not love Romeo has an incentive to send such a signal. Only if Juliet can choose between a noisy "I love you" signal and a noisy "I don't love you" signal does it become credible that she is honest. Thus, a single new signal observed cannot convince of Juliet's honesty. Romeo must know that Juliet is using a *system* of noisy signals. Still, in a pooling equilibrium, evolutionary drift could predispose Rome to respond to two noisy signals in an appropriate way, after which Juliet would find it a best response to send them.

3.4 Imprecise signaling

Under the pressure of the local tourist industry, a weather forecaster would like to systematically predicts higher temperatures than will truly happen at the seaside – unless the temperature will be perfect anyway. When the temperature will be 20°C, the weather forecaster would like the public to believe the temperature is 22°C, when it will 22°C he would like them to believe it will be 24°C, and when it will be 24°C he would like them to believe it is indeed 24°C. How can the forecaster's prediction still be credible? The forecaster can achieve credibility by being imprecise. Given that she does not want the public's expectations to deviate too widely from the truth, if she makes her signals imprecise, she gives herself an incentive to send these signals in a truthful way. In this case, she could predict that it will be 20°C if this is indeed true, and predict that it will be 22°C or more in all other cases, where the latter leads to an expectation by the public that it will be 24°C. Given this response by the public to her imprecise signal, the forecaster does not want to predict that it will be at least 22°C when it will be only 20°C. In general, the principle of pragmatic inference applied here is that the sender is so imprecise that the incentive to cheat is no longer present.

Conflict-of-interest imprecise signaling model. The model reflecting this story is a simplified version of Crawford and Sobel (1981). There are three states of the world *L*, *M* and *R*, and three responses with the same label. In every state, the receiver prefers to take the action with the same label. When only knowing that state *R* does not occur, the receiver prefers response *L*. In state *R*, the sender's most preferred response is *M*, but she still prefers that the receiver adopts response *R* rather than *L*. In both states *M* and *L*, she prefers him to take action *L*. As shown in Proposition 9, the conflict of interest is resolved if the sender sends the same signal in states *M* and *L*. This is the only separating equilibrium.

Proposition 9. Consider the conflict-of-interest imprecise signaling model. Assume that $U_Z(M|R) > U_Z(R|R) > U_Z(L|R)$, $U_Z(L|S) > U_Z(M|S)$, $U_Z(L|S) > U_Z(R|S)$ for $S = L, M$. Let $\frac{[U_H(L|L) - U_H(M|L)]}{[U_H(M|M) - U_H(L|M)]} > \frac{\pi_M}{\pi_L}$, meaning that a receiver who knows that state *R* does not occur prefers to take action *L*. Then in the only separating Nash equilibrium, the sender does the same in state *M* and *L* (send a signal or not send a signal).

Proof:

There is no equilibrium where *Z* honestly reveals all states, since otherwise *Z* in state *R* would send a message inducing action *M*, and *Z* in state *M* would send a message inducing action *L*.

There is no equilibrium where Z pools states M and R. This is because Z in state M would then send the signal inducing L. As $\frac{\left[U_H(L|L)-U_H(M|L)\right]}{\left[U_H(M|M)-U_H(L|M)\right]}>\frac{\pi_M}{\pi_L}$, when Z pools M and L, H does L. It follows that Z in state L and M do not want to send the signal inducing R. At the same type, Z in R does not want to send the pooled signal, as Z in R prefers action R to action L.
QED

Checking whether the unique separating equilibrium is selected against equilibrium refinement arguments, we need not look at iterated elimination of strictly dominated strategies, or at the intuitive criterion. This is because signals are assumed cheap here, so that the receiver cannot conclude that they could only have been sent by certain types.

Proposition 10. Consider the conflict-of-interest imprecise signaling model.
(i) The *pooling equilibria* are not neologism proof and cannot be attained from any separating equilibrium.
(ii) The *separating equilibrium* is neologism proof, and *can* be attained from pooling equilibria.
Proof:
(i) Pooling equilibria are not neologism proof, whatever the form they take. If R is the receiver's best response in the pooling equilibrium, consider a neologism "R does not occur". Following equation (1), by $U_Z(R|R)>U_Z(L|R)$, $\frac{\left[U_H(L|L)-U_H(M|L)\right]}{\left[U_H(M|M)-U_H(L|M)\right]}>\frac{\pi_M}{\pi_L}$, this cannot have been sent by the type R sender. By $U_Z(R|M)<U_Z(L|M)$, $U_Z(R|L)<U_Z(L|L)$, the type M and type L sender has an incentive to send such a signal. If L is the best response in the pooling equilibrium, consider a neologism of the form "state R occurs". By $U_Z(R|M)<U_Z(L|M)$, $U_Z(R|L)<U_Z(L|L)$, and $U_H(R|R)>U_H(L|R)$, equation (1) tells that type L and M senders do not have an incentive to send such a signal. By $U_Z(R|R)>U_Z(L|R)$, equation (2) tells that the type R sender does have an incentive to send such a message. If M is the best response in the pooling equilibrium, consider a neologism of the form "state M or L occurs". By $U_Z(M|R)>U_Z(L|R)$ and $\frac{\left[U_H(L|L)-U_H(M|L)\right]}{\left[U_H(M|M)-U_H(L|M)\right]}>\frac{\pi_M}{\pi_L}$, the type R sender does not have any incentive to send a message inviting response L rather than response M. By $U_Z(M|M)<U_Z(L|M)$, $U_Z(M|L)<U_Z(L|L)$, the type M and L senders do prefer to send such a signal.
Starting from the separating equilibrium, whatever the response of the receiver when no signal is received, it is not in the sender´s interest to stop sending signals.
(ii) Starting from the separating equilibrium, a neologism "do M". could only have been sent by the player in state R, so that the receiver will still respond with R to such a signal. If in the separating equilibrium, the sender sends a signal "do L or M", a neologism "do L".could not have been sent by type R, so that the receiver will do L when receiving it. But this does not give M or L any incentive to send this signal, as they already obtain L. – By the previous arguments, a neologism do "do L or R" does not destabilize the separating equilibrium, whatever the receiver´s response to it.

In the pooling equilibrium, it is a best response for the receiver to respond with *L* to an out-of-equilibrium signal, making it a best response for the sender to employ his best response in the separating equilibrium.

QED

Intuitively, as there is only a single separating equilibrium, it is not a problem to reach this from the pooling equilibrium, both by means of attainability and by means of neologisms.

4. Games with common interests

In the conflict-of interest models of Section 3, the driving principle is that *Z* is biased towards always reporting the same state of the world. If this state of the world actually occurs, the only manner in which *Z* can still convince *H* of this is by making his signal costly in such a manner that another type of *Z* would never want to put up with. *H* infers from the costly signal that this must have come from one type of *Z*. This can happen by revealing information that only a sender in this state could reveal, by making one's signal costly to send in terms of the direct cost of sending it, or indirectly in terms of its noisiness or imprecision.

In the current section, we treat common-interest models. For costly signaling, noisy signaling and imprecise signals, the intuition is each time that it is efficient for sender and receiver to incur any unavoidable costs of signaling only in states where this cost has the least impact. Thus, a costly signal should only be sent in infrequent states, so that the cost is incurred as little as possible. Similarly, the cost of noisiness or imprecision should be left for states where this has less impact, again because these states are infrequent or because mistakes are less costly in these states. In our common-interest direct signaling model, the intuition is different. There, it is efficient to trust that revelations by the sender that are as such inconclusive (e.g., a single piece of good news) still tell something about the overall state of the world (e.g., overall good news).

4.1 Direct signaling

Returning to the dictator example of Section 3.1, we now assume that the dictator and the inspector have common interests. Thus, if she were able to do so, the dictator would reveal to the inspector her status with respect to both chemical and nuclear weapons. Unfortunately, due to costs she is only able to reveal either evidence on whether or not she has nuclear weapons, or evidence on whether or not she has chemical weapons. Objectively, if e.g., she reveals that she does not have nuclear weapons, this is unconvincing to the receiver. Yet, sender and receiver can coordinate on an equilibrium where the dictator reveals whatever type of weapons she has if she has only one type of them, reveals any type if she has both of them, and reveals herself not to have one of the types if she has neither of them. While objectively speaking evidence that one of the types is not in her possession does not prove that she has no weapons at all, given that the dictator has good intentions, the receiver still infers that there are no weapons. The principle of pragmatic inference applied here is that as the signaler has the same interests, an as such inconclusive piece of bad news is interpreted as overall bad news, and an as such inconclusive piece of good news is interpreted as overall good news.

Direct signaling common-interests model with two cues underlying the states of the world. The model is identical to the two-cue model in Section 3.1, except that sender and receiver have common interests. The sender still is only able to reveal a single cue. A separating equilibrium now exists where the sender only reveals one of the cues to be "0" (no

nuclear weapons, or no chemical weapons) if in fact *both* cues are "0", and does not reveal anything otherwise. This is in spite of the fact that literally speaking, a single "0" revealed does not give the receiver sufficient direct information that state *L* occurs ("the dictator does not have weapons of mass destruction"). In terms of the dictator example, the dictator's evidence that he does not have one type of weapons is interpreted as evidence that he has neither type of weapons. Further, a pooling equilibrium exists, where the form of the receiver's response depends on the parameters.

Proposition 11. Consider the direct signaling common-interests model with two cues underlying the states of the world.

(i) If $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} > \frac{(1-x)^2}{x(2-x)}$, in the pooling equilibrium action *M* is always taken.
If $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} < \frac{(1-x)^2}{x(2-x)}$, in the pooling equilibrium action *L* is always taken.

(ii) A separating equilibrium exists where the sender reveals a "0" if both of the cues have a value of 0, and does not reveal anything otherwise. A separating equilibrium also exists where the sender reveals a "1" if at least one of the cues has a value of 1, and nothing otherwise. Finally, a separating equilibrium exists where the sender reveals a "0" only if both cues have value of zero, and reveals a "1" if at least one of two values is 1.

Proof:
For the proof of (i), see the proof of Proposition 3. (ii): if the receiver believes that a single revealed "0" means that both cues have a value of 0, then given common interests the sender follows the candidate equilibrium strategy. This in turn makes it a best response for the receiver to follow the candidate equilibrium strategy. The same applies to the other separating equilibria.
QED

The results in Proposition 11 show a fundamental difference with the conflict-of-interest model. With a conflict of interest, in a separating equilibrium, the sender reports only the value of a *specific* cue and never the information on the other cue in order to avoid giving the impression that she only reports information that is to her advantage. With common interests, the sender reports the value of any cue that happens to be representative for the overall state of the world. We now investigate whether equilibrium selection eliminates the pooling equilibrium in favor of such a separating equilibrium in the case of common interests.

Proposition 12. Consider the direct signaling common-interests model with two cues underlying the states of the world.

(i) If $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} > \frac{(1-x)^2}{x(2-x)}$, the pooling equilibrium survives iterated elimination of strictly dominated strategies, meets the intuitive criterion, but cannot be attained from any separating equilibrium. Every separating equilibrium survives the intuitive criterion and is attainable from the pooling equilibrium, but does not destabilize the pooling equilibrium.

(ii) If $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} < \frac{(1-x)^2}{x(2-x)}$, the pooling equilibrium does not survive iterated elimination of strictly dominated strategies, does not meet the intuitive criterion, and cannot be attained from any separating equilibrium. Every separating equilibrium survives

the intuitive criterion and is attainable from the pooling equilibrium, but only the separating equilibrium where the sender reveals a "0" if at least one of the cues is a zero, and does not reveal anything otherwise, destabilizes the pooling equilibrium.

Proof:

In general, any separating equilibrium is attainable from any pooling equilibrium, as in a pooling equilibrium it is always a best response to respond in the proper way to an out-of-equilibrium revealed cue.

(i) It is strictly dominated for a receiver to whom a "1" is revealed to do anything but *M*. But this does not lead to elimination of the pooling equilibrium by iterated elimination of strictly dominated strategies, as the receiver already does *M* in the pooling equilibrium. A sender who observes at least one 1 and reveals a "0" or a "1" is not strictly better off in equilibrium than with the best he can achieve by not revealing a "0" or a "1" (namely equally well obtain response *M*). A sender who observes two values of 0 is better off by revealing a "0" if this leads the receiver to do *L*. It follows that there is no type for which revelation of a cue is equilibrium dominated, so that the pooling equilibrium meets the intuitive criterion.

(ii) It is strictly dominated for a receiver to whom a "1" is revealed to do anything but *M*. Given that *L* is always done in the pooling equilibrium, the sender deviates and reveals a "1" if at least one of his cues has a value of 1. Thus, the pooling equilibrium is eliminated by iterated elimination of strictly dominated strategies in favor of the separating equilibrium. Further, it is impossible for the sender who does not observe 1s to reveal a 1. It follows that by equation (5), the sender who observes at least a 1 is better off by revealing a "1", so that the pooling equilibrium does not meet the intuitive criterion.

QED

As shown by Proposition 12, in terms of the dictator example, if in the pooling equilibrium the receiver concludes that the dictator does *not* have weapons of mass destruction, then this equilibrium is not stable because the receiver knows that a cue that the dictator does have such weapons can only be sent by a dictator who indeed has them. If in the pooling equilibrium the receiver instead concludes that the dictator *does* have weapons of mass destruction, then a revealed cue that the dictator does not have nuclear weapons, or does not have chemical weapons, does not destabilize the pooling equilibrium by iterated elimination of strictly dominated strategies or by the intuitive criterion, as it is not the case that only a dictator that does not have any weapons at all can send such a signal or has an incentive to send such a signal. This suggests that the principle of pragmatic inference proposed here, saying that the limited information that is revealed should be considered as representative for all information, only applies if the pooling equilibrium has a particular form. Still, by evolutionary drift the receiver may get predisposed to interpret an out-of-equilibrium revealed cue that the sender does not have weapons of mass destruction, and a cue that she does have them, in the appropriate way. A separating equilibrium then still evolves from the pooling equilibrium, so that the principle still applies.

4.2 Costly signaling

Consider two drivers driving in opposite directions on the same road, and assume that they have common interests. Driver 1 can see whether or not there is a speed control ahead for driver 2. Most of the time, there is no speed control ahead. The only signal available for driver 1 to send to driver 2 is to flash her headlights. But how should driver 2 interpret driver 1 flashing her headlights? As flashing one's headlights requires some effort, the most efficient outcome is that flashing one's headlights means that there is a speed camera ahead. In this way, the effort of sending the signal is incurred as infrequently as possible. The principle of pragmatic inference applied here is that an unexpected signal means that something out of the ordinary is going on. (Un)marked states receive an (un)marked expression. This principle goes back to Horn and Zipf, and was game-theoretically treated among others by Parikh (1991, 2000, 20001) and Van Rooij (2004).

Costly signaling common-interests model. There are two states of the world, *L* and *M*, and only actions *L* or *M* can be taken. In an essential feature of this game, we assume that state *L* is less frequent than state *M*, i.e. $\pi_L < \pi_M$. The sender can choose to send any signal from a set of equally costly signals. As shown in Proposition 13, in terms of the driver example, there exists an efficient separating equilibrium where the first driver flashes her headlights only if there is a speed control, as well as an inefficient equilibrium where the first driver flashes her headlights only if there is no speed control ahead. Further, depending on the parameters, in the pooling equilibrium either the second driver drives slow or fast.

Proposition 13. Consider the costly signaling common-interests model. Then an efficient separating equilibrium exists where *Z* sends the costly signal only in state *L*, along with an inefficient separating equilibrium where the *Z* sends a costly signal only in state *M*. In the pooling equilibrium, *H* does *L* when $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} \leq \frac{\pi_L}{\pi_M}$, and does *M* when $\frac{\left[U_H(M|M)-U_H(L|M)\right]}{\left[U_H(L|L)-U_H(M|L)\right]} \geq \frac{\pi_L}{\pi_M}$. The Pareto-efficient equilibrium is the one where a signal gets sent in the infrequent state *L*.

Proof:

If $\pi_M U_H(M|M)+\pi_L(M|L) \geq \pi_M U_H(L|M)+\pi_L(L|L)$, then a pooling equilibrium where *M* is always done exists, if $\pi_M U_H(M|M)+\pi_L(M|L) \leq \pi_M U_H(L|M)+\pi_L(L|L)$, then a pooling equilibrium where *L* is always done exists. If the sender sends a signal in state *L* (*M*), then it is a best response for the receiver to take action *L* (*M*) when a signal is received, and to take action *M* (*L*) when a signal is not received. This response by the receiver in turn makes the specified sender strategy a best response. For the receiver, it does not matter which separating equilibrium is played. The sender is better off if the signal is only sent infrequently.

QED

We now again look at equilibrium selection arguments, to investigate whether the principle that a signal should be interpreted as referring to an out of the ordinary state indeed is predicted to apply. The arguments here are due to Van Rooij (2008) and De Jaegher (2008). In comparison with the costly-signaling conflict-of-interest model of Section 3.2, it should be noted that there are no strictly dominated strategies anymore, where certain types can never send certain signals. Instead, only the principle of equilibrium domination can be applied, where the comparison is made to what *Z* of a certain type could have obtained by playing

another equilibrium. For this reason, iterated elimination of strictly dominated strategies is irrelevant.

Proposition 14. Consider the costly signaling conflict-of-interest model.

Let $\frac{\left[U_H\left(M|M\right)-U_H\left(L|M\right)\right]}{\left[U_H\left(L|L\right)-U_H\left(M|L\right)\right]} > \frac{\pi_L}{\pi_M}$.

(i) *Pooling equilibria* are not neologism proof (against efficient separating equilibria), do not meet the intuitive criterion, do not destabilize any separating equilibria and cannot be attained from them.

(ii) The *inefficient separating equilibrium* is neologism proof, meets the intuitive criterion, but does not destabilize pooling equilibria or the efficient separating equilibrium, and cannot be attained from them.

(iii) The *efficient separating equilibrium* is neologism proof, meets the intuitive criterion, destabilizes pooling equilibria but not efficient separating equilibria, and can be attained from pooling equilibria but not from the inefficient separating equilibrium.

Let $\frac{\left[U_H\left(M|M\right)-U_H\left(L|M\right)\right]}{\left[U_H\left(L|L\right)-U_H\left(M|L\right)\right]} < \frac{\pi_L}{\pi_M}$. Then:

(iv) *Pooling equilibria*: not neologism proof (against inefficient separating equilibria), do not meet intuitive criterion, do not destabilize any separating equilibria and cannot be attained from them.

(v) *Inefficient separating equilibria*: neologism proof, meet intuitive criterion, destabilize pooling equilibria but not efficient separating equilibria, can be attained from pooling equilibria but not from efficient separating equilibria.

(vi) *Efficient separating equilibria*: neologism proof, meet intuitive criterion, but do not destabilize pooling equilibria or inefficient separating equilibria, and cannot be attained from them.

Proof:

In general, any separating equilibrium is neologism proof and meets the intuitive criterion because no type can benefit from sending an out-of-equilibrium signal. Each sender type is already obtaining the preferred response in a separating equilibrium. As each signal is costly, at best the sender obtains exactly the same payoff by sending an out-of-equilibrium signal. A Pareto inefficient (efficient) equilibrium separating equilibrium is not attainable from a Pareto efficient (inefficient) separating equilibrium, as the receiver in the Pareto inefficient (efficient separating equilibrium does *L* (*M*) when not receiving a signal.

(i) If a costly signal already has a common meaning "do *L*" and if the receiver follows the advice contained in the signal, type *M* does not want to send it by $U_Z\left(M|M\right) > U_Z\left(L|M\right)$ (equation (1)), but type *L* does by $U_Z\left(L|L\right) > U_Z\left(M|L\right)$ (equation (2)), so that the pooling equilibrium is not neologism proof. Given $U_Z\left(M|M\right) > U_Z\left(L|M\right)$ it is further equilibrium dominated for type *M* to send a signal (equation (4)). It follows by equation (5) that the pooling equilibrium does not meet the intuitive criterion and is destabilized by the efficient separating equilibrium.

(ii) In any pooling equilibrium in this case, the receiver does *M* when not receiving a signal. The sender's strategy of the inefficient separating equilibrium is not a best response to this, so that the inefficient separating equilibrium is not attainable from pooling equilibria.

(iii) In any pooling equilibrium in this case, it is a best response for the receiver to do *M* when not receiving a signal and *L* when receiving a signal. The sender's strategy of the efficient

separating equilibrium is a best response to this receiver strategy, so that the efficient separating equilibrium is attainable from pooling equilibria.

(iv) If a costly signal already has a common meaning "do *M*" and if the receiver follows the advice contained in the signal, type *L* does not want to send it by $U_Z(L|L) > U_Z(M|L)$ (equation (1)), but type *M* does by $U_Z(M|M) > U_Z(L|M)$ (equation (2)), so that the pooling equilibrium is not neologism proof. Given $U_Z(L|L) > U_Z(M|L)$ it is further equilibrium dominated for type *L* to send a signal (equation (4)). It follows by equation (5) that the pooling equilibrium does not meet the intuitive criterion and is destabilized by the inefficient separating equilibrium.

(v) In any pooling equilibrium in this case, it is a best response for the receiver to do *L* when not receiving a signal and *M* when receiving a signal. The sender's strategy of the inefficient separating equilibrium is a best response to this, so that the inefficient separating equilibrium is attainable from pooling equilibria.

(vi) In any pooling equilibrium in this case, it is a best response for the receiver to do *L* when not receiving a signal. The sender's strategy of the efficient separating equilibrium is not a best response to this receiver strategy, so that the efficient separating equilibrium is not attainable from pooling equilibria.

QED

(dit keer te specifiek in termen van voorbeeld)

In terms of the driver example, Proposition 14 suggests that the proposed principle of pragmatic inference only applies if in the pooling equilibrium, the second driver drives fast. This occurs e.g., if speed controls are infrequent. Even though it is not the case that only a driver who sees a speed camera ever wants to flash her headlights (no strict dominance), it is still the case that only a driver who sees a speed camera has an incentive to deviate from the pooling equilibrium and incur the cost of flashing her headlights. Yet, if in the pooling equilibrium the driver drives slow even though speed controls are infrequent, e.g., because traffic fines are very high, then by the same principle the second driver would flash her headlights most of the time, when no speed control is ahead. Thus, from the perspective of neologism proofness and of the intuitive criterion, the principle is rather to look at what sender has an incentive to deviate from the pooling equilibrium by incurring the cost of sending a signal. Nevertheless, a recent experiment (De Jaegher, Rosenkranz and Weitzel, 2008) shows that even if in terms of the example in the pooling equilibrium the second driver drives slow, the efficient separating equilibrium is still played. This suggests that the efficiency of the Pareto-efficient equilibrium creates a focal point.

4.3 Noise

Some weeks ago, a conference organizer invited an academic as a speaker at his conference, and the academic agreed to come. The night before the conference day, the organizer realizes that he has not heard from the academic since she agreed to come weeks ago. Even though the organizer and the academic have common interests, the organizer wonders how to interpret the academic's silence. Does this simply mean that she will show up as agreed, and would she only have sent a new message to cancel her talk due to unforeseen circumstances? Or does her silence mean that she forgot about the conference, and is it the case that if she would not have forgotten, he would have received a re-confirmation from her right before the conference?

An additional complication is that no matter whether the academic sends confirmation or cancellation messages, the message may get lost. But given that any message may get lost, the organizer may consider in what circumstances it is most costly for a message to get lost. If it is costly to schedule the academic when she does not show up (so that the audience might be left waiting) but not costly to have the academic present without her being able to hold her talk, it is efficient that a confirmation message is sent. In this way, the audience is never left waiting. If it is costly that the academic show up and finds that she cannot hold her talk but not costly for the audience that her talk is scheduled (because there are parallel sessions), it is efficient that a cancellation message is sent. In this way, the academic can always hold her talk if she is present. In general, as silence does not lead to mistakes (an e-mail will not be received if none was sent), no message should be sent in the case where it is important not to make mistakes. Sender and receiver then follow a pragmatic rule of being being unambiguous when it is important.

Common-interests noisy signaling model. In the simplest model reflecting this story, there are only two states of the world, *M* and *L*, and two actions with the same label. Both sender and receiver want the label of the receiver's action to be the same as the label of the state of the world. The sender can choose a signal from a set of signals. The signals, which may or may not be part of a rich common language, all get lost with the same probability. We only consider equilibria where the sender either sends a message or does not.

Proposition 15. Consider the common-interests noisy signaling model. Let *M* be *H*'s best response in the pooling equilibrium. Then an efficient separating equilibrium exists where *Z* sends a noisy only in state *L*, along with an inefficient separating equilibrium where *Z* sends a noisy signal only in state *M*.

Proof:

$$\pi_M U_H(M|M) + \pi_L\left[\mu(S|S)U_H(L|L) + \mu(0|S)U_H(M|L)\right]$$

$$>$$

$$\pi_M\left[\mu(S|S)U_H(M|M) + \mu(0|S)U_H(L|M)\right] + \pi_L U_H(L|L)$$

$$\Leftrightarrow$$

$$\pi_M\left[U_H(M|M) - U_H(L|M)\right] > \pi_L\left[U_H(L|L) - U_H(M|L)\right]$$

QED

In Proposition 16, we now check to what extent equilibrium refinement arguments elect the Pareto-efficient separating equilibrium. We again leave out iterated elimination of strictly dominated strategies, and the intuitive criterion, as these both only work if there is a direct cost attached to sending a signal.

Proposition 16. Consider the common-interests noisy signaling model. Let *M* be *H*'s best response in the pooling equilibrium. Then:

(i) The *pooling equilibria* are not neologism proof against the efficient separating equilibrium, and cannot be attained from any of the separating equilibria.

(ii) The *inefficient separating equilibrium* is neologism proof, but cannot be attained from pooling equilibria or from efficient separating equilibria.

(iii) The *efficient separating equilibria* is neologism proof, cannot be attained from inefficient separating equilibria, but can be attained from pooling equilibria.

Proof:

In general, pooling equilibria cannot be attained from a separating equilibrium, and one type of separating equilibrium cannot be attained from the other. With respect to the pooling

equilibrium, drift cannot cause the sender to prefer to stop sending signals. With respect to the other type of separating equilibrium, drift cannot cause the sender to prefer to reverse the circumstances where a signal is sent and not sent. All separating equilibria are neologism proof against the pooling equilibria, simply because an out-of-equilibrium neologism cannot lead to the undoing of separation. Equilibria were two messages are sent were not considered as part of the model

Given that *M* is played in the pooling equilibrium, by equation (1), a sender in state *M* would *not* prefer to send a signal with commonly-known meaning "do *L*" if the receiver would follow this advice. By equation (2), the sender in state *L* does weakly prefer to send such a signal. However, by equation (1), a sender in state *M* only has a weak interest to send a signal with commonly-known meaning "do *M*", given that *M* is already done in the pooling equilibrium. It follows that the pooling equilibrium is not neologism proof against the Pareto-efficient separating equilibrium.

Given that *M* is done in any pooling equilibrium, drift cannot cause the receiver to do *M* only when receiving a signal and to do *L* otherwise, but can cause the receiver to do *L* only when receiving a signal and *M* otherwise.
QED

Proposition 16 only shows that neologism proofness and attainability both predict that the efficient separating equilibrium will be played in case *M* is the best response in the pooling equilibrium. But this is a general result: by simply reversing all labels, it can be seen that the result also applies when *L* is played in the pooling equilibrium. In terms of the example at the beginning of this section, if the key consideration is that the audience is never left waiting, then an organizer who knows that the academic does not send any signals will not schedule the academic. For the same reason, it is efficient that the academic sends a confirmation. In this way, the audience is never left waiting (while she may show up at the conference to find out that her talk was canceled because her confirmation got lost). But given that her talk is canceled if she is known not communicate, the only thing that she can learn is to send a confirmation. While in neologism proofness and attainability, Pareto-efficiency is not the driving force, they happen to lead to Pareto-efficiency in this case, so that the pragmatic rule telling to be unambiguous when it is important would always be applied.

4.4 Imprecision

A doping agency performs doping tests on athletes, and the results of its tests can either be positive, negative, or inconclusive. The sports federation to which it reports prefers to suspend athletes with a positive test, to give athletes with an inconclusive test a warning, and to leave athletes with a negative test free. Unfortunately, the doping agency can only give the sports federation a yes answer or a no answer on whether the athlete uses doping, so that inevitably one of the two possible judgements of the doping agency needs to cover two states. The sports federation and the doping agency both hate to let dishonest athletes escape suspension. For this reason, they both prefer that athletes who test positive and inconclusive are suspended, and that only athletes with a negative test go free. When inferring what failure to report a positive test result means, the sports federation may conclude that the agency is following a pragmatic rule to "be precise when it is important", i.e. to make sure that negative tests are only inferred for athletes who are certainly not positive.

A related idea is found in a recent paper by Crémer et al. (2007), who show that it is efficient for firms to use precise, fine-tuned jargon in states of the world with which they are confronted frequently, and imprecise language in states of the world that are less frequent.

The argument there, however, is one of design rather than coordination between a sender and a receiver. A much more sophisticated model is found in Jäger et al. (forthcoming), who study a common-interest signaling game with continuous states of the world defined over multiple dimensions (e.g., location on a plane) and with a finite number of signals. While most of the analysis concerns uniform distributions (all states are equally likely) and while the focus is on the form of equilibrium partitions (Voronoi tessellations), the authors also pay attention to non-uniform distributions, where in an example it is shown to be efficient that types with a larger mass in the distribution function have smaller partitions, whereas types with less mass have larger partitions. Yet, there are also inefficient equilibria, and the focus in our analysis is whether equilibrium refinements select the efficient equilibria.

Common-interest imprecise signaling model. There are three states of the world *L*, *M* and *R*. Each player prefers that in each state, the action with the same label as the state is taken. Action *R* is taken in the pooling equilibrium, as well as when *R* and *M* are pooled. When *M* and *L* are pooled, action *M* is taken.

Proposition 17. Consider the three-state three-action game with a set $\mathbb{S}$ ($|\mathbb{S}| \geq 1$) of noiseless, costless signals, where there are common interests. Let *Z* be constrained to use only a *single* signal in any separating equilibrium. Let $U_Z(L|L) > U_Z(M|L) > U_Z(R|L)$, $U_Z(R|R) > U_Z(M|R) > U_Z(L|R)$, and $U_Z(M|M) > U_Z(R|M) > U_Z(L|M)$. Further, let $\frac{[U_H(L|L) - U_H(M|L)]}{[U_H(M|M) - U_H(L|M)]} < \frac{\pi_M}{\pi_L}$, $\frac{[U_R(M|M) - U_R(R|M)]}{[U_R(R|R) - U_R(M|R)]} < \frac{\pi_R}{\pi_M}$, $\frac{[U_H(M|M) - U_H(R|M)]}{[U_H(L|L) - U_H(M|L)]} < \frac{\pi_L}{\pi_M}$, and let $\pi_L[U_H(x|L) - U_H(R|L)] + \pi_M[U_H(x|M) - U_H(R|M)] < \pi_R[U_H(R|R) - U_H(x|R)]$ for $x = M, L$. Then in the pooling equilibrium, *H* does *R*. In the inefficient separating equilibrium, *Z* pools states *M* and *L* (leading *H* to do *M*). In the efficient separating equilibrium, *Z* pools states *M* and *R* (leading *H* to do *R*).

Proof:

(i) The pooling equilibrium takes this form by the fact that $\pi_L[U_H(x|L) - U_H(R|L)] + \pi_M[U_H(x|M) - U_H(R|M)] < \pi_R[U_H(R|R) - U_H(x|R)]$ for $x = M, L$.

(ii) In the first type of equilibrium, *M* and *L* get pooled. By $\frac{[U_H(L|L) - U_H(M|L)]}{[U_H(M|M) - U_H(L|M)]} < \frac{\pi_M}{\pi_L}$, *H* does *M* when getting a pooled signal. Given that $U_Z(M|L) > U_Z(R|L)$, $U_Z(R|R) > U_Z(M|R)$, and $U_Z(M|M) > U_Z(R|M)$, *Z* prefers to tell the truth.

(iii) In the second type of equilibrium, *R* and *M* get pooled. By $\frac{[U_R(M|M) - U_R(R|M)]}{[U_R(R|R) - U_R(M|R)]} < \frac{\pi_R}{\pi_M}$, *H* does *R* when getting a pooled signal. Given that $U_Z(L|L) > U_Z(R|L)$, $U_Z(R|R) > U_Z(L|R)$, and $U_Z(R|M) > U_Z(L|M)$, *Z* prefers to tell the truth.

(iv) $\pi_L U_H(M|L) + \pi_M U_H(M|M) + \pi_R U_H(R|R) < \pi_L U_H(L|L) + \pi_M U_H(R|M) + \pi_R U_H(R|R)$ iff $\frac{[U_H(M|M) - U_H(R|M)]}{[U_H(L|L) - U_H(M|L)]} < \frac{\pi_L}{\pi_M}$.

QED

We now look at equilibrium selection arguments in this case. The pooling equilibrium can be eliminated, but both types of separating equilibria (the efficient as well as the inefficient

one) seem equally likely to come forward. Neologism proofness does not work in selecting the efficient separating equilibrium because both types of equilibria make the sender better off compared to the pooling equilibrium. (We do not treat the neologism proofness of separating equilibria. If an additional signal is available, obviously a separating equilibrium will evolve where the three states are separated. The point of the exercise is that only a limited number of signals is available.) Additionally, the attainability argument does not work anymore either. Clearly, from a situation where in the pooling equilibrium R is always done, it may evolve that a signal is sent only in state L, or is sent both in state L and state M. Thus, while the argument for the efficient equilibrium seems intuitive, only the focal point argument selects the efficient equilibrium. In terms of the doping agency example, a pragmatic rule that the doping agency should only be precise when it is important to be precise (i.e. when the athlete certainly did not take doping) only works if it somehow creates a focal point.

Proposition 18. Consider the game in Proposition 17.

(i) *Pooling equilibria*: not neologism proof against both efficient and inefficient separating equilibria, but cannot be attained from any of the separating equilibria.

(ii) *Inefficient separating equilibria*: neologism proof, can be attained from pooling equilibria.

(iii) *Efficient separating equilibria*: neologism proof, can be attained from pooling equilibria.

Proof:

(i) Consider any pooling equilibrium. First, let H receive a signal from a common language saying that either M or L occurs. If the signal is followed up, following equation (1), Z in R does not have any strong incentive to send such a signal, Z in M or L do (equation (2)). Second, let H receive a signal from a common language saying that L occurs. If the signal is followed up, given that $U_Z(R|M) > U_Z(L|M)$ and $U_Z(R|R) > U_Z(L|R)$, neither Z in R nor Z in M have an incentive to send such a signal, but Z in L does have an incentive to send such a signal.

(ii) In the pooling equilibrium, it is a best response to the receiver to adopt towards an out-of-equilibrium signal the strategy of the inefficient or of the efficient separating equilibria. It follows that both these equilibria are attainable from the pooling equilibrium.

QED

5. Discussion

As the paper has shown, game-theoretic pragmatics can be applied not only to costly signals, but also to noisy, imprecise, and to direct signals. In part, we have shown this by reinterpreting or adapting the existing literature, but we have also shown in simple examples how some gaps in the literature concerning in particular direct and noisy signals could be filled. This was done in terms of very basic models, suggesting that additional insights can be gained by studying more sophisticated models. Moreover, the models we treat have each time considered revelatory aspects of signals, and their cost, noisiness and imprecision separately. Yet, signals typically may have all these aspects at the same time. Thus, it may be insightful to also consider models where signals have several of these aspects at the same time.

As our analysis shows, standard equilibrium refinements such as neologism proofness or the intuitive criterion do not always predict that efficient separating equilibria are more likely to

be played than inefficient ones. Indeed, in some instances these arguments predict the inefficient separating equilibrium to be played. The same applies for evolutionary arguments. This creates a tension with the focal-point argument, saying that the efficiency of a separating equilibrium itself creates a focal point, so that the efficient separating equilibrium is played. These contradictory predictions are interesting from the viewpoint of laboratory experiments, where the games can be replicated with pecuniary payoffs, to test which of the equilibrium selection arguments has cutting ground.

Finally, we are aware that under the influence of similar trends within economics, game-theoretical pragmatics is starting to move into behavioral game theory (see e.g., Franke, 2011). While we find this a sensible approach, as shown in this paper we have not explored all insights we can gain from rational models yet. One could say that we do not go through that phase, and go straight to the behavioral models, without lingering in the rational ones. Yet, we believe that the rational models are needed as a basis, before we can start to construct behavioral models. We need the benchmark of rational models to compare our behavioral models against.

References

Binmore, K., and Samuelson, L. (1999) Evolutionary drift and equilibrium selection. *Review of Economic Studies*, 66: 363-93.

Blume, A., Board, O.J. and Kawamura, K. (2007) Noisy talk. *Theoretical Economics*, 2: 395-440.

Blume, A. and Board, O. (2008) Intentional vagueness. working paper, University of Pittsburgh.

Cho, I.K., and Kreps, D. (1987) Signaling Games and Stable Equilibria. *Quarterly Journal of Economics* 102, 179–222.

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

Crémer, J, Garicano, L, and Prat, A. (2007) Language and the theory of the firm. *Quarterly Journal of Economics* 122, 373-407.

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

De Jaegher, K. (2003b), A game-theoretic rationale for vagueness. *Linguistics and Philosophy* 26, 637-659.

De Jaegher, K. (2008) The evolution of Horn's rule. *Journal of Economic Methodology* 15, 275-284.

De Jaegher, K., Rosenkranz, S. and Weitzel, U. (2008) Economic laboratory experiment on Horn's rule. Working paper 08-27, Utrecht University.

De Jaegher, K. and Rooij, R. van (2011) Strategic vagueness and appropriate contexts, in: *Language, Games and Evolution – Trends in Current Research on Language and Game Theory*, A. Benz, C. Ebert, G. Jäger and R. van Rooij (eds.), Springer Verlag, Heidelberg, pp. 40-59.

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

Franke, M. and Jager, Tikitu de (2011) Now that you mention it – awareness dynamics in discourse and decisions, in: *Language, Games and Evolution – Trends in Current Research on Language and Game Theory*, A. Benz, C. Ebert, G. Jäger and R. van Rooij (eds.), Springer Verlag, Heidelberg, pp. 60-91.

Glazer, J. and Rubinstein, A. (2001) Debates and decisions: on a rationale of argumentation rules. *Games and Economic Behavior* 36, 158-173.

Grice, H.P. (1967). Logic and Conversation. William James Lectures, Harvard University, reprinted in (1989) *Studies in the Way of Words*, Cambridge, MA: Harvard University Press, pp. 22–40.

Horn, L. (1984). Towards a New Taxonomy of Pragmatic Inference: Q-Based and R-Based Implicature, in *Meaning, Form, and Use in Context: Linguistic Applications*, ed. D. Schiffrin, Washington, DC: Georgetown University Press, pp.11–42.

Jäger, G., Metzger, L.P. and Riedel, F. Voronoi languages: Equilibria in cheap-talk games with high-dimensional types and few signals. Games and Economic Behavior, forthcoming.

Johnstone, R.A. (1997) The evolution of animal signals. In: Krebs, J.R., Davies, N.B. (Eds.), *Behavioural Ecology*. Blackwell, Oxford, pp. 155–178.

Milgrom, P. (1981) Good news and bad news: representation theorems and applications. *Bell Journal of Economics* 13, 380-391.

Myerson, R.G. (1991) *Game Theory – Analysis of Conflict*. Harvard University Press, Cambridge.

Parikh, P. (1991) Communication and strategic inference. *Linguistics and Philosophy* 14: 473-513.

Parikh, P. (2000) Communication, meaning, and interpretation. *Linguistics and Philosophy* 23: 185-212.

Parikh, P. (2001) *The Use of Language*, Stanford, California: CSLI Publications.

Pawlowitsch, C. (2008) Why evolution does not always lead to an optimal signaling system. *Games and Economic Behavior* 63: 203-226.

Rooij, R. van. (2004), Signaling games select Horn strategies. *Linguistics and Philosophy* 27, 493–527.

Rooij, R. van (2008) Games and Quantity Implicatures. *Journal of Economic Methodology* 15, 261-274.

Sobel, J. (1993) Evolutionary stability and efficiency. *Economics Letters* 42, 301-312.

Schelling (1960) *The Analysis of Conflict*. Harvard University Press, Cambridge.

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

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