Bidirectional Optimization from Reasoning and Learning in Games

Michael Franke & Gerhard Jäger
Department of Linguistics
University of Tübingen

September 19, 2011

Abstract

We reopen the investigation into the formal and conceptual relationship between bidirectional optimality theory (Blutner, 1998, 2000) and game theory. Unlike a likeminded previous endeavor by Dekker and van Rooij (2000), we consider signaling games not strategic games, and seek to ground bidirectional optimization once in a model of rational step-by-step reasoning and once in a model of reinforcement learning. We give sufficient conditions for equivalence of bidirectional optimality and the former, and show based on numerical simulations that bidirectional optimization may be thought of as a process of reinforcement learning with *lateral inhibition*.

1 Bidirectional Optimization

Optimality Theory (OT) is an abstract framework used in many linguistic subdisciplines to study the mapping between different levels of representation. Its original applications fell mostly into phonology (see Prince and Smolensky, 1993, and much subsequent work), but the framework, due to its abstract explanatory versatility, was soon after applied to other domains of grammar as well. Bidirectional optimality theory (BIOT), as conceived by Blutner (1998, 2000), advanced standard unidirectional OT for applications to interface problems within and between syntax, semantics and pragmatics, where it is used to study the relation between a level of forms and meanings (cf. Hendriks and de Hoop, 2001; Blutner and Zeevat, 2004).

Let $\mathcal{F}$ be a set of forms and $\mathcal{T}$ be a set of meanings.[1] An OT-*system* $\langle \mathcal{G}, \preceq \rangle$ comprises a relation $\mathcal{G} \subseteq \mathcal{T} \times \mathcal{F}$ between these two levels, called the *generator*, and a preference ordering $\preceq \subseteq \mathcal{G} \times \mathcal{G}$. This relation is a total preorder, i.e., it is reflexive, transitive, and total. For $g, g' \in \text{Gen}$ we write $g \prec g'$ if $g \preceq g'$ and $g' \npreceq g$. If $g \prec g'$, then g is strictly better (more harmonic) than g'. We demand that $\prec$ be *well-founded*, i.e., that each subset of $\mathcal{G}$ has a minimal element with respect to $\prec$. In particular, this ensures that $\prec$

[1] Letters $\mathcal{T}$ and $\mathcal{F}$ are mnemonic for types and forms. These terms are motivated by the tradition of signaling games (see below).

is acyclic. To exclude degenerate cases, we demand that $Dom(\mathcal{G}) = \mathcal{T}$ and $Rg(\mathcal{G}) = \mathcal{F}$. We write $\mathcal{G}(t, f)$ for $\langle t, f\rangle \in \mathcal{G}$, as well as $\mathcal{G}(t) = \{f \in \mathcal{F} \mid \mathcal{G}(t, f)\}$ and similarly for $\mathcal{G}(f)$.

Standard unidirectional OT has a simple protocol to identify the optimal pairings between inputs and outputs: $\langle t, f\rangle$ is *(unidirectionally) optimal* if $\mathcal{G}(t, f)$, and there is no f' with $f' \prec_t f$. (We write $t_1 \prec_f t_2$ whenever $\langle t_1, f\rangle \prec \langle t_2, f\rangle$, and likewise for $\langle f_1 \prec_t f_2\rangle$.) In words, the optimal output for a given input t is simply the minimal element(s) with respect to $\prec_t$. Most work in OT revolves around identifying an ordered set of constraints that implicitly define $\preceq$ as a lexicographic ordering over constraint violation profiles. The focus of this paper is different though. We would like to scrutinize the notion of optimality, not the coming-about of $\preceq$, with special emphasis on the notion of *bidirectional optimality* central to BIOT.

Bidirectional optimization is based on the intuitive insight that optimal communication is a dyadic process that involves both a speaker that has to find an optimal way to express her communicative intentions, and a listener that has to find an optimal interpretation for the signal that he observed. What makes things even more complicated is the fact that these two optimization procedures have to be intertwined. A message is only optimal for a given content if the listener is able to recover the content from the message. Likewise, an interpretation of a message is only optimal if this is really the interpretation that the speaker had in mind. This recursive reference of speaker optimization to hearer optimization and vice versa was captured by Blutner (1998, 2000) in the following way (modulo slightly different notation and terminology):

- f is *speaker-optimal* for t iff $\mathcal{G}(t, f)$ and there is no f' with $f' \prec_t f$ such that t is hearer-optimal for f'.
- t is *hearer-optimal* for f iff $\mathcal{G}(t, f)$ and there is no t' with $t' \prec_f t$ such that f is speaker-optimal for t'.
- $\langle t, f\rangle$ is *bidirectionally optimal* iff f is speaker-optimal for t and t is hearer-optimal for f.

So both speaker and hearer try to find a map that is minimal with respect to $\prec$. However, they both restrict their search space to the candidates that are optimal for the other interlocutor. The fact that $\prec$ is well-founded ensures though that this iterated change of perspective eventually terminates. Consequently, the notion of bidirectional optimality is in fact formally well-defined, as shown in Jäger (2002). There it is also shown that this definition is equivalent to a somewhat simpler formulation:

Definition 1.1 (Bidirectional Optimality). A pair $\langle t, f\rangle \in \mathcal{G}$ is *bidirectionally optimal* iff

- there is no $t' \prec_f t$ such that $\langle t', f\rangle$ is bidirectionally optimal, and
- there is no $f' \prec_t f$ such that $\langle t, f'\rangle$ is bidirectionally optimal.

Again, well-foundedness is crucial here because it guarantees that the elements of $\mathcal{G}$ are arranged in a linear hierarchy. A pair $\langle t, f\rangle$ is *level 0* if it has no predecessors with respect to $\prec$. A pair $\langle t, f\rangle$ is *level (α)* if all predecessors of $\langle t, f\rangle$ have level lower than α, and α is the smallest number with this property. If $\prec$ is well-founded, every pair will have a well-defined level, which is always some, possibly transfinite, ordinal number. If a certain pair $\langle t, f\rangle$ is of level k and you want to determine whether it is bidirectionally optimal, you only have to check alternative candidates of lower levels. Optimality can thus be defined in a cumulative fashion, starting with candidates at level 0 and successively proceeding through higher levels until $\mathcal{G}$ is exhausted:

Definition 1.2 (Bidirectional Optimality, Cumulative). The set O of *bidirectionally optimal* pairs is defined as:

$$\begin{aligned}
O_0 &= \emptyset \\
O_{\alpha+1} &= O_\alpha \cup \{\langle t, f\rangle \in \mathcal{G} - Dom(O_\alpha) \times Rg(O_\alpha) \mid \\
&\qquad \forall t' \prec_f t : t' \in Dom(O_\alpha) \wedge \forall f' \prec_t f : f' \in Rg(O_\alpha)\} \\
O_\beta &= \bigcup_{\alpha<\beta} O_\alpha \ (\beta \text{ is a limit ordinal}) \\
O &= \bigcup O_\alpha\,.
\end{aligned}$$

This definition is (modulo change in notation) identical to the corresponding definition provided by Jäger (2002), who proves it to be equivalent to the definition above.

The transfinite recursion is bounded by the order type of $\mathcal{G}$, i.e. by the highest type of any element of $\mathcal{G}$. In particular, iff $\mathcal{G}$ is finite, the third clause of the cumulative definition is actually superfluous, and the definition can be turned straightforwardly into the following algorithm:

```
$O = \emptyset$
while ($\mathcal{G} - Dom(O) \times Rg(O) \neq \emptyset$) do
  $O \rightarrow O \cup \{\langle t, f\rangle \mid \forall t' \prec_f t : t' \in Dom(O) \wedge \forall f' \prec_t f : f \in Rg(O)\}$
end while
return $O$
```

Let us apply this to the example that is depicted in Figure 1. Here we have three meanings and three forms, with the preference relation as indicted by the arrows pointing to an adjacent *better* pair. (Transitive arrows are omitted, non-optimal pairs that occur in the generator are marked with ∘ and bidirectionally optimal pairs are marked by •). This OT-system is a generalized case of M-implicature, or "Horn's division of pragmatic labor", according to which unmarked forms (1a) preferably pair with unmarked meanings (1b), and marked forms (2a) preferably pair with marked meanings (2b) (c.f. Horn, 1984; Levinson, 2000).

(1) a. Black Bart killed the sheriff.

b. $\rightsquigarrow$ Black Bart killed the sheriff in a stereotypical way.

(2) a. Black Bart caused the sheriff to die.

b. $\rightsquigarrow$ Black Bart killed the sheriff in a non-stereotypical way.

Solving this OT-system with the BIOT-algorithm, $\langle t_1, f_1\rangle$ is added to O in the first iteration of the *while* loop, as this is the only pair that has no predecessors in either dimension.

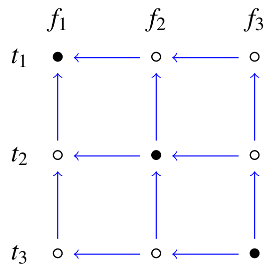


Figure 1: Generalized M-implicature

This entails that $\langle t_1, f_2\rangle, \langle t_1, f_3\rangle, \langle t_2, f_1\rangle$ and $\langle t_3, f_1\rangle$ are *blocked*. As they share one component with an optimal pair but are not optimal themselves, they are not considered anymore in future iterations and come out as non-optimal. In the second iteration, only $\langle t_2, f_2\rangle, \langle t_2, f_3\rangle, \langle t_3, f_2\rangle$, and $\langle t_3, f_3\rangle$ are still in play. Among them, only $\langle t_2, f_2\rangle$ is added to O, while $\langle t_2, f_3\rangle$ and $\langle t_3, f_2\rangle$ are blocked. In the third iteration $\langle t_3, f_3\rangle$ is the only remaining candidate, which is thus added to O. After the third iteration, the stop condition of the loop is fulfilled and the algorithm returns $O = \{\langle t_1, f_1\rangle, \langle t_2, f_2\rangle, \langle t_3, f_3\rangle\}$.

So far we have only given abstract definitions with a minimum of conceptual motivation. But how may we conceptualize bidirectional optimality? Surprisingly, proponents of BIOT are not unanimous about this issue. Some consider bidirectional optimization as a diachronic, evolutionary force that shapes language use over longer intervalls of time (Blutner and Zeevat, 2004, 2008). Others consider bidirectional optimality a model of pragmatic reasoning competence (cf. Hendriks et al., 2007, chapter 5 for discussion of different possible positions). It is this issue that this paper speaks to. We ask how the abstract black box called "bidirectional optimality" could be grounded in more tangible concepts, such as beliefs, preferences, rationality and learning.

2 Optimization from Iterated Best Responses

The notion of bidirectional optimization is motivated by communicators' *strategic* considerations: a speaker who tries to find the optimal form for a given meaning will take the hearer's optimization process into account and vice versa. Such decision making in interactive situations that takes the preferences of other decision makers into account is the domain of *game theory*. Since game theory is well-understood, both notionally, as well as mathematically, the hope is justified that matching BIOT to as close a formal game-theoretic analog as possible will help delineate how bidirectional optimization could be interpreted.

Signaling Games for OT-Systems. Dekker and van Rooij (2000) explore the parallel between BIOT and *strategic games*. This, however, is conceptually implausible for language use, as it assumes that speakers and hearers make choices *simultaneously* and *unconditionally*, which in turn means that a speaker would choose what to say regardless of what she is trying to express, and that a hearer would choose an interpretation before or independently of hearing a message uttered. Therefore, we have argued elsewhere (Franke, 2009) that OT-systems should better be translated into *signaling games* (Lewis, 1969), where first the speaker chooses a message for a given meaning that she wants to express, and the hearer subsequently tries to guess at this meaning.

For any given OT-system we can construct a signaling game as follows. A signaling game has two players, a sender S and a receiver R. The sender has some private information, her *type* t. The set of types can be identified with OT's $\mathcal{T}$. Types are assigned to senders according to a fixed *prior* probability distribution p^* that is common knowledge between the players. For simplicity, we will assume throughout that p^* is uniform: $p^*(t) = p^*(t')$ for all t, t'. For each type $\mathcal{T}$, S has a choice among a set of messages $\mathcal{G}(t)$ that she can transmit to R; in other words, in translating an OT-system into a signaling game we assume that only forms that occur in the generator for a given meaning are feasible signals for the corresponding type in the corresponding signaling game (c.f. the games of partial information of Parikh, 1987). Next, after observing a message, R has a choice among a set of actions. We only consider the case where R tries to guess S's type. Thus, upon observing f, R chooses among $\mathcal{G}(f)$; again, the strategy space is restricted by $\mathcal{G}$. Taken together, a single round of play is characterized by a sequence type-message-type, called a *history*. Both players have preferences over histories, which are captured by utility functions u_s and u_r (S's and R's utility functions respectively) that map triples $\langle t_1, f, t_2\rangle \in \mathcal{T} \times \mathcal{F} \times \mathcal{T}$ into a numerical measure of success (usually a real number, but see below). The question left to ask is how to properly translate the preference relation of a given OT-system. Several options suggest themselves, but only few are ultimately sensible. In the second half of this paper, we implement OT-preferences as initial *behavioral dispositions* of learning agents. Another obvious and reasonable option for a rationalistic approach is to model OT-preferences in terms of the players' utility functions. But how exactly?

It is common to assume that at the heart of non-degenerate meaningful talk exchange lies an overarching interest in successful communication that is shared by all participants. This is what we will assume too. But additionally we will also assume that players may have further secondary preferences due to their possibly diverging notions of communicative efficiency encoded in the OT-system's preference ordering. This can be cast into a utility function that implements a *lexicographic ordering* in that it returns a pair of numbers for each history (which we will treat as a vector in later computation):

$$u_s(t_1, f, t_2) = \langle \delta(t_1, t_2),\ -c(t_1, f)\rangle \qquad u_r(t_1, f, t_2) = \langle \delta(t_1, t_2),\ -c(t_2, f)\rangle\ . \tag{1}$$

Here, δ is the Kronecker function that returns 1 if its two arguments are identical, and 0 otherwise. This models the interlocutors' shared interest in successful communication. Subordinate to that, $c : \mathcal{T} \times \mathcal{F} \to \mathbb{R}$ is a *cost function* which corresponds to the OT-system's preference ordering in the obvious way:

$$\langle t_1, f_1\rangle \preceq \langle t_2, f_2\rangle \quad \text{iff} \quad c(t_1, f_1) \leq c(t_2, f_2)\,. \tag{2}$$

When comparing these utilities, costs are treated as *nominal*, i.e., they only ever play a role when comparing options that are equally successful in communication. This is ensured by the following postulate on lexicographic ordering of vectors:

$$\langle (x_1, y_1) \rangle < \langle (x_2, y_2) \rangle \quad \text{iff} \quad x_1 < x_2 \vee (x_1 = x_2 \wedge y_1 < y_2) \tag{3}$$

Strategies and Best Responses. A (pure) strategy for S is a function from $\mathcal{T}$ into $\mathcal{F}$. Likewise, a strategy for R is a function from $\mathcal{F}$ into $\mathcal{T}$. It is a standard assumption of rationalistic game theory that players will make their decisions in a way that maximizes their expected utility. The expected utility, however, depends on the opponent's strategy, about which the decision maker may have only partial information. This is captured in the notion of a *behavioral strategy*. A behavioral strategy for S is a function σ that assigns to each type a probability distribution over messages:

$$\forall t : \sigma(\cdot|t) \in \Delta(\mathcal{G}(t)) \, .$$

σ is R's belief about S's behavior. Likewise, S's belief about R's behavior is captured by a behavioral strategy ρ assigning to each form a probability distribution over types:

$$\forall f : \rho(\cdot|f) \in \Delta(\mathcal{G}(f)) \, .$$

A *best response* to a behavioral strategy is a strategy that chooses in each situation an option that maximizes the expected utility. A rational player will always play according to a best response to his assumptions about the other player. When calculating his best response, R makes use of Bayes' rule:

$$\sigma(t|f) = \frac{\sigma(f|t)p^*(t)}{\sum_{t'} \sigma(f|t')p^*(t')} \, .$$

The best responses of a player are thus as follows:[2]

$$BR_r(f;\sigma) = \begin{cases} \arg_t \max \sum_{t'} \sigma(t'|f)\, u_r(t', f, t) & \text{if } \max_t \sigma(f|t) > 0 \\ \mathcal{G}(f) & \text{otherwise.} \end{cases}$$

$$BR_s(t;\rho) = \arg_f \max \sum_{t'} \rho(t'|f)\, u_s(t, f, t') \, .$$

Step-By-Step Reasoning in Pragmatics. In rationalistic game theory it is commonly assumed that rational players will settle in a *Nash equilibrium*. A Nash equilibrium is a strategy profile where each player plays a best response to the opponents' strategy. However, research in behavioral game theory has shown that human reasoners often do not confirm the predictions of equilibrium in experimental settings. Rather, human reasoners seem to employ step-by-step reasoning heuristics when making a decision in a strategic situation (cf. Selten, 1998).

[2] Recall that $u_{s,r}(t, f, t')$ is a vector for which the max-operation is defined on a lexicographic ordering. Also, it should be mentioned that the second condition in the definition of BR_r does not necessarily follow from a definition of best responses in terms of *posterior beliefs*. We take the liberty of glossing over alternative, more technical derivations here. For the purposes of this paper suffice it to give the motivation that for *surprise messages*, i.e, signals f for which $\max_t \sigma(f|t) = 0$, we assume that the receiver is *completely indifferent* which interpretation to chose.

For these and other reasons, various game-theoretic approaches to communication have explored solution concepts that substitute equilibrium notions for iterative reasoning protocols (cf. Stalnaker, 2006; Benz and van Rooij, 2007; Jäger, 2008; Franke, 2009). The basic idea that is common to these approaches is this: player *A* starts her reasoning process with the provisional assumption that the other player, *B*, follows some non-rational default strategy; based on this assumption, *A* chooses her best possible strategy, i.e., the strategy that maximizes *A*'s payoff; however, *B* might anticipate *A*'s reasoning step and not play the default strategy but rather a best response to *A*'s strategy, and so on. This procedure may be iterated arbitrarily many times and may or may not lead to equilibrium eventually.

Such *iterated best response models* (IBR models), as we will call them here, show a striking similarity with bidirectional optimization, especially in the latter's iterative, algorithmic guise. Still, the two approaches are not entirely equivalent. In the remainder of this section we will compare BiOT with IBR. We show similarities but also consider the major conceptual difference, namely a different stance towards the "distributional information" contained in gaps of the generator. Subsequently, we give a set of sufficient conditions under which the approaches are nonetheless guaranteed to coincide.

An Iterated Best Response Model. We assume that IBR reasoning starts with a naïve receiver strategy ρ_0 that chooses any possible interpretation with equal probability for each message.[3] We then recursively define a sequence of behavioral strategies by best responses to the opponent's behavior at the previous step. If there are more than one optimal options in a given situation, we apply the *principle of insufficient reason* and assume that the corresponding behavioral strategy assigns equal probability to all these optimal choices. (We represent uniform distributions simply by their support set.)

Definition 2.1 (IBR sequence).

$$\rho_0(t|f) = |\mathcal{G}(f)|^{-1}$$

$$\sigma_n(f|t) = \begin{cases} |BR_s(t;\rho_n)|^{-1} & \text{if } f \in BR_s(t;\rho_n) \\ 0 & \text{otherwise.} \end{cases}$$

$$\rho_{n+1}(t|f) = \begin{cases} |BR_r(f;\sigma_n)|^{-1} & \text{if } t \in BR_r(f;\sigma_n) \\ 0 & \text{otherwise.} \end{cases}$$

If the game is finite, this sequence will eventually reach a fixed point which is an equilibrium (see Franke, 2011, for a proof). Players who reason to depth ω, but not necessarily only those, will play according to these fixed-point equilibrium strategies.

It is instructive to apply this model to the example in Figure 1. We do not need specific numerical values, as the assumption of nominal costs and the constraint in Equation (2) are sufficient to calculate the IBR sequence. The result is given in Figure 2. (The matrix notation for behavioral strategies is to be understood in the sense that each argument in the leftmost column is mapped to the uniform distribution over the values in the rightmost column.) $\langle \rho_3, \sigma_3 \rangle$ is a fixed point, i.e., for all $n > 3$, $\langle \rho_n, \sigma_n \rangle =$

[3] An IBR sequence can also be defined starting from a naïve sender strategy. To keep things simple, we focus here only on the sequence that starts with a naïve receiver.

$$\rho_0 = \begin{bmatrix} f_1 & \rightarrow & t_1, t_2, t_3 \\ f_2 & \rightarrow & t_1, t_2, t_3 \\ f_3 & \rightarrow & t_1, t_2, t_3 \end{bmatrix} \qquad \sigma_0 = \begin{bmatrix} t_1 & \rightarrow & f_1 \\ t_2 & \rightarrow & f_1 \\ t_3 & \rightarrow & f_1 \end{bmatrix}$$

$$\rho_1 = \begin{bmatrix} f_1 & \rightarrow & t_1 \\ f_2 & \rightarrow & t_1, t_2, t_3 \\ f_3 & \rightarrow & t_1, t_2, t_3 \end{bmatrix} \qquad \sigma_1 = \begin{bmatrix} t_1 & \rightarrow & f_1 \\ t_2 & \rightarrow & f_2 \\ t_3 & \rightarrow & f_2 \end{bmatrix}$$

$$\rho_2 = \begin{bmatrix} f_1 & \rightarrow & t_1 \\ f_2 & \rightarrow & t_2 \\ f_3 & \rightarrow & t_1, t_2, t_3 \end{bmatrix} \qquad \sigma_2 = \begin{bmatrix} t_1 & \rightarrow & f_1 \\ t_2 & \rightarrow & f_2 \\ t_3 & \rightarrow & f_3 \end{bmatrix}$$

$$\rho_3 = \begin{bmatrix} f_1 & \rightarrow & t_1 \\ f_2 & \rightarrow & t_2 \\ f_3 & \rightarrow & t_3 \end{bmatrix} \qquad \sigma_3 = \begin{bmatrix} t_1 & \rightarrow & f_1 \\ t_2 & \rightarrow & f_2 \\ t_3 & \rightarrow & f_3 \end{bmatrix}$$

Figure 2: IBR reasoning for the generalized M-implicature

$\langle \rho_3, \sigma_3 \rangle$. So the IBR sequence eventually selects the same mapping between types and forms that BIOT selects. But the correspondence is even closer: the set of pairs $\mathcal{I}_n = \{\langle t, f \rangle \mid \rho_n(t|f) = 1\}$ is actually identical to the set O_n from Definition 1.2 for all n.

This is not a coincidence. The cumulative definition of BIOT has in fact a strong affinity to the IBR logic. At each stage of bidirectional optimization, only those types are considered that are not yet optimally paired to any message. Likewise, only those messages are considered that are not yet paired to any type. Among the Cartesian product of these sets, those pairs are identified as optimal that are not *blocked* by a better alternative in either dimension.

The IBR sequence operates similarly. Suppose we have reached a stage n with a behavioral receiver strategy ρ_n and a corresponding set $\mathcal{I}_n$. If a type t is in the domain of $\mathcal{I}_n$, this means that there is at least a message f which is mapped to t with probability 1 under ρ_n. A rational sender that plays a best response to ρ_n will map t to such a message. If, however, $t \notin Dom(\mathcal{I}_n)$, this means that there is no message that will induce the receiver's action t with probability 1. In the present example, the best the sender can do here is to send a *surprise message*, i.e., a message that is not in the range of $\mathcal{I}_n$. These are messages to which the receiver assigned probability 0 at the previous stage, and which are thus mapped to the uniform distributions over all possible types. Choosing such a message f will yield a posterior probability of $|\mathcal{G}(f)|^{-1}$, which is better than 0. Furthermore, a rational sender will choose the cheapest message with this property. So upon observing a message f that is a surprise message for ρ_n, the receiver at level ρ_{n+1} will infer that this was sent by a type that is un-inducible under ρ_n, and he will map f to the least complex of these types. So $\mathcal{I}_{n+1}$ will contain all the pairs from $\mathcal{I}$, plus all pairings of uninducible types with surprise messages which are outperformed by alternatives along either dimension.

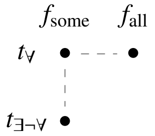
Major Difference: Quantity Reasoning. We hasten to add that this superficial likeness of BiOT and IBR does not hold in full generality. The most obvious conceptual difference is that rationalistic IBR always implements a particular kind of "quantity reasoning", i.e., reasoning about the comparative logical strength of alternative signals as covered by Grice (1975)'s Maxim of Quantity. On the other hand, BiOT does not do so unless specified by additional constraints, and this may lead to diverging predictions in cases where the generator $\mathcal{G}$ is a proper subset of $\mathcal{T} \times \mathcal{F}$.

For instance, consider as a standard case of a *scalar implicature* the inference from an utterance of (3a) to the truth of (3c) based on the idea that an informed cooperative speaker would have uttered the semantically stronger (3b) had it been true.

(3) a. Some of Kiki's friends are metalheads.

b. All of Kiki's friends are metalheads.

c. Some but not all of Kiki's friends are metalheads.



In its simplest form, an OT-system for this case would be the one given in example (3) where there are no preferences at all between pairs (indicated by dashed gray lines in the diagram). It is easy to see that, as the diagram also shows, all pairs in the generator are bidirectionally optimal in this case. This is *not* the solution of IBR, because, even without any secondary payoffs, IBR factors in the fact that hearing f_{some} makes $t_{\exists\neg\forall}$ twice as likely as state $t_{\forall}$. The IBR fixed point therefore selects the natural bijective mapping $\{\langle t_{\exists\neg\forall}, f_{\text{some}}\rangle, \langle t_{\forall}, f_{\text{all}}\rangle\}$, unlike BiOT.

Unfortunately, there is no way to throw out the bathwater ("quantity reasoning" of this kind) *without* the baby (a standard notion of rationality). If we are still determined to pursue the parallel between BiOT and rationalistic IBR, we need to restrict our attention to OT-systems which also implement a constraint Quantity that is ranked highest:[4]

$$|\mathcal{G}(f_1)| > |\mathcal{G}(f_2)| \quad \rightarrow \quad \forall t : f_2 \prec_t f_1\,. \tag{4}$$

A Cumulative Reformulation of IBR. If we consider only OT-systems where Quantity is the highest-ranked constraint, and where additionally the following condition on costs (see Equation (2)), which we call NoTies, holds:[5]

$$\forall g, g' \in \mathcal{G} \quad : \quad c(g) = c(g') \quad \rightarrow \quad g = g' \tag{5}$$

[4] Indeed that is what we find, for instance, in Aloni (2007)'s OT-approach to conversational implicatures, where a quantity-constraint is only outranked by a truth-fullness. Similarly, a quantity-constraint was also assumed in BiOT's forefather, the theory presented by Blutner (1998).

[5] This condition is not unnatural, and frequently made in both OT, as well as game theory: often we would like results to be stable under slight random shocks of the input parameters, but this is only possible if nothing is ranked exactly alike.

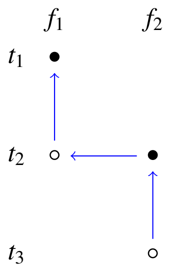


Figure 3: Problematic example 1

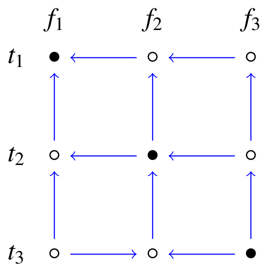


Figure 4: Problematic example 2

we can offer a cumulative reformulation of IBR in terms of the OT-system's preference ordering as follows:

$$\begin{aligned}
\mathcal{I}_0 &= \emptyset \\
\mathcal{I}_{n+1} &= \mathcal{I}_n \cup \{\langle t, f\rangle \in \mathcal{G} - Dom(\mathcal{I}_n) \times Rg(\mathcal{I}_n) \mid \\
&\qquad \forall t' \prec_f t (t' \notin Dom(\mathcal{I}_n) \rightarrow \exists f' \prec_{t'} f : f' \notin Rg(\mathcal{I}_n)) \wedge \\
&\qquad \forall f' \prec_t f : f' \in Rg(\mathcal{I}_n)\} \\
\mathcal{I} &= \bigcup_{n \in \omega} \mathcal{I}_n .
\end{aligned}$$

Theorem 2.2. If an OT-system satisfies Equations (4), (5) and $\forall f \in \mathcal{F} : |\mathcal{G}(f)| > 1$, then for any signaling game constructed as described (with uniform priors and nominal costs) we have $\mathcal{I}_n(t, f)$ iff $\rho_n(t|f) = 1$ for all n.

A proof is given in Appendix A.

Further Differences. The cumulative formulation of IBR looks *nearly* like the cumulative formulation of BIOT. But sadly, outer appearance notwithstanding, this reformulation brings further differences to light. Here, we will discuss briefly two contrived examples of OT-systems for which predictions differ. To the best of our knowledge, these examples, and others like them, have no immediate empirical bearing: unlike the case of quantity reasoning, we cannot presently conceive of any linguistic use for which the noted formal differences would be relevant. Still, these examples serve to highlight further conceptual differences between IBR and BIOT. For completeness' sake, we pursue a purely formal investigation, and give a constraint that rules out aberrant cases like these and thus entail equivalence nonetheless.

The example in Figure 3 does satisfy the Quantity constraint, but still the predictions of IBR and BIOT differ. The former associates f_2 and t_3 in its fixed point, but the bidirectionally optimal solution associates f_2 with t_2.[6] Another problematic case

[6] Examples of this kind could be ruled out by an additional constraint that mirrors Quantity also for the speaker, i.e., a constraint of the form: $|\mathcal{G}(t_1)| > |\mathcal{G}(t_2)| \rightarrow \forall f : t_2 \prec_f t_1$. This, however, would not help with the subsequent example either.

is exemplified by the OT-system in Figure 4 (which differs from the example in Figure 1 only insofar as the preferences between $\langle t_3, f_1\rangle$ and $t_3, f_2\rangle$ are reversed). The BIOT solution for this case is $\mathcal{O} = \{\langle t_1, f_1\rangle, \langle t_2, f_2\rangle, \langle t_3, f_3\rangle\}$, while IBR predicts the solution $\mathcal{I} = \{\langle t_1, f_1\rangle, \langle t_2, f_3\rangle, \langle t_3, f_2\rangle\}$.

Together, both examples illustrate an interesting difference between IBR and BIOT. While IBR is flexible and retains ordering information that also involves pairs that are blocked by the BIOT-algorithm, the workings of the latter are rather adequately described by a "first-come-first-serve" heuristic. Nonetheless, equivalence between $\mathcal{O}$ and $\mathcal{I}$ can be ensured if we assume that the following condition, which we could call SQUARING, holds for all t_1, t_2, f_1 and f_2:[7]

$$\Big(t_1 \prec_{f_1} t_2 \;\wedge\; f_1 \prec_{t_2} f_2\Big) \;\rightarrow\; \Big(\mathcal{G}(t_1, f_2) \;\wedge\; f_1 \prec_{t_1} f_2 \;\wedge\; t_1 \prec_{f_2} t_2\Big)\,. \tag{6}$$

Theorem 2.3. If an OT-system meets conditions QUANTITY, NOTIES and SQUARING from Equations (4), (5) and (6), then any signaling game properly constructed to our purpose will satisfy $\mathcal{I}_n = \mathcal{O}_n$ for all n, and $\mathcal{I} = \mathcal{O}$.

Proof. We prove the first statement by induction over n. The case $n = 0$ is trivial. So let $\mathcal{O}_n = \mathcal{I}_n$. Suppose $\mathcal{O}_{n+1}(t, f)$. Then it follows directly that $\mathcal{I}_{n+1}(t, f)$. Now suppose $\mathcal{I}_{n+1}(t, f)$. Then $\forall f' \prec_t f : f' \in Rg(\mathcal{O}_n)$. Suppose $\exists t' \prec_f t$ such that $t' \notin Dom(\mathcal{O}_n)$. It follows from the definition of IBR that there is an $f' \prec_{t'} f : f' \notin Rg(\mathcal{O}_n)$. Hence $f' \prec_t f$, according to the constraint from Equation (6). This is a contradiction, i.e., $\forall t' \prec_f t : t' \in Dom(\mathcal{O}_n)$, and this entails that $\mathcal{O}_{n+1}(t, f)$. As $\mathcal{G}$ is finite, the second claim follows immediately. □

Summary. To sum up briefly, it is possible to recast BIOT in rationalistic terms as IBR reasoning but there are also significant conceptual differences between these approaches. The main difference is that "quantity reasoning" is a direct spin-off of game-theoretic IBR, while BIOT needs additional constraints for this. On top of that, as the examples in Figures 3 and 4 showed, a theory of rational agency has trouble accommodating in general especially two characteristic features of bidirectional optimization: (i) its *fast-and-frugal* association mechanism, and (ii) its *lateral blocking behavior*. Consequently and despite superficial parallels and likeness, it is reasonable to conclude that optimality does not square well with a too sophisticated agent architecture. We would therefore also like to consider next a model of optimization through try-and-error learning in which agents are not sophisticated at all.

3 Bidirectional Optimality & Reinforcement Learning

In this section, we consider an interpretation of bidirectional optimization as the outcome of *reinforcement learning*, where agents grope towards optimality gradually by myopic trial-and-error, without forming beliefs about opponent behavior, let alone rationally responding to behavioral beliefs. We hypothesize that a particular process of

[7] SQUARING follows, for instance, from the (stronger) assumption that (i) $\mathcal{G} = \mathcal{T} \times \mathcal{F}$, and (ii) all OT constraints are markedness constraints.

reinforcement is conceptually closely related to the selection and blocking behavior of the BIOT-algorithm. Here is a first plausibility argument in support of this idea.

Suppose that we translate a given OT-system into a signaling game in the way we did above, except that now we assume that the agents' utility functions are simply defined by the desire to communicate successfully: $u_{s,r}(t_1, f, t_2) = \delta(t_1, t_2)$. Further assume that an agent's behavior is not defined via beliefs and best responses, but given by a probabilistic choice function, the origins and dynamics of which are not subject to the agents' introspection and rational deliberation. Let us assume that the speaker's and the hearer's preferences, as defined by OT's $\preceq$, give the initial inclinations of players to act in the sense that, for instance, a sender in state t would be more inclined to use a message f_i rather than a message f_j just in case $f_\mathrm{i} \prec_t f_\mathrm{j}$. Starting with these initial probabilistic strategies, players play the game repeatedly and whenever they achieve successful communication the choices that lead to success are more likely to be repeated in consecutive trials. After many steps of this trial-and-error learning with reinforcement of successes, signaling behavior may or may not emerge.

How would that relate to the BIOT-algorithm? Consider a pair $\langle t, f \rangle$ which is optimal after the first round of the BIOT-algorithm. By initial set-up, successful communication is more likely to occur with $\langle t, f \rangle$ than with competing pairs $\langle t', f \rangle$ or $\langle t, f' \rangle$. A probabilistic variant of BIOT's selection-&-blocking is then simply a concomitant of reinforcement: if the receiver has successfully chosen interpretation t for message f, he is more likely to chose it again in later rounds of play, and also, by rescaling, less likely to chose t' after f in the future. A parallel argument applies to the sender, of course.

The remainder of this section therefore tries to assess the hypothesis that the BIOT-algorithm describes *the most likely path of reinforcement learning*, and that consequently the bidirectionally optimal pairs describe the most likely convergence point of this dynamics. We present results of numerical simulations that support this hypothesis, at least for certain types of reinforcement learning. To properly introduce the relevant notions, the following section first introduces some background on reinforcement learning.

3.1 Reinforcement Learning & Pólya Urns

A very accessible and simple approach to reinforcement learning can be formulated in terms of so-called *Pólya urns*. A probabilistic choice function of an agent can be modelled by an urn that contains balls of different colors, one color for each action the agent may choose in the given situation. The proportions of balls of a given color in the urn give the probabilities with which the agent chooses corresponding actions. A concrete action choice is modelled by a random draw from the relevant urn. To model the effects of learning by reinforcement, the proportion of the balls may subsequently be amended, depending on how successful the performed action was. In the simplest case success is binary: each action is either successful or not. If we only consider *positive reinforcement*, then if the chosen action was not successful, we simply return the drawn ball of the corresponding color, leaving the proportions of colors in the urn unchanged; but if the action was successful, we return the drawn ball, together with $\alpha > 0$ other balls of the same color, thus increasing the likelihood that this color will be drawn on consecutive trials with this urn. The number α is a parameter that regulates how much reinforcement takes place when an action is successful. If we also consider *negative reinforcement*, then we subtract a number of balls $\beta \geq 0$ from an urn in case the previous action was *not* successful. We do so unless this would remove all balls of a given type from an urn. The number β is a parameter that regulates how severly failure is punished, and we consider positive values for β, but also the case $\beta = 0$, where there is no negative reinforcement. We call α the *bonus* and β the *malus*. It is obvious that the learning process depends on these parameters, but it also depends on the initial number of balls in the urns.

We will also study the effect of so-called *lateral inhibition*, captured by a third parameter, the *suppressor* γ. However, lateral inhibition can be explained much more easily after having introduced urn-based reinforcement learning for signaling games.

3.2 Reinforcement Learning for Signaling Games

Several contributors have recently turned towards studying Pólya-urn models of reinforcement learning as a learning dynamics for the evolution of signaling (Barrett, 2009; Skyrms, 2010; Mühlenbernd, 2010). To model reinforcement learning in games, each choice point of each agent is associated with an urn. For the sender, this means that there is one urn $\mathcal{U}_t$ for each state t of the signaling game. Any $\mathcal{U}_t$ is filled with balls corresponding one-to-one with each message. We assume that whenever $f \notin \mathcal{G}(t)$, then $\mathcal{U}_t$ does not contain any balls for f. This way a set of sender urns, one for each state, models a legitimate behavioral sender strategy for a signaling game derived from an OT-system as described above. Similarly, for the receiver with his urns $\left\{\mathcal{U}_f\right\}_{f \in \mathcal{F}}$.

We assume that sender and receiver start from an initial probabilistic strategy that implements a given OT's preference relation. Obviously, since this gives no quantitative information, a map from the OT-ordering to an order-preserving probabilistic strategy is one-to-many. As a kind of neutral approach, we suggest that the proportions between probabilities of elements that are adjacent in the relevant OT-ordering should be constant. In oder to implement this idea, say, for the sender, we look at the rank $\text{Rk}(f)$ for each message f given the ordering $\prec_t$ (with rank 1 for the minimal elements in the order and higher ranks for "better" elements), and fill the urn $\mathcal{U}_t$ with $\exp(\lambda, \text{Rk}(f))$ balls for each message f if $f \in \mathcal{G}(t)$ and no balls otherwise. Here, the *ranking factor* λ is a parameter that decides how much probability difference occurs between differently ranked messages. The receiver's urns are initially filled analogously. For example, the players' urns in the signaling game derived for the OT-system in Figure 3 are initially filled as follows for ranking factor $\lambda = 10$:

	f_1	f_2
$\mathcal{U}_{t_1}$	10 balls	0 balls
$\mathcal{U}_{t_2}$	100 balls	10 balls
$\mathcal{U}_{t_3}$	0 balls	10 balls

	t_1	t_2	t_3
$\mathcal{U}_{f_1}$	100 balls	10 balls	0 balls
$\mathcal{U}_{f_2}$	0 balls	100 balls	10 balls

A single learning step consists of (i) the sender drawing a ball from each of her urns, and (ii) the receiver drawing a ball from each of the urns whose messages where selected in (i). This corresponds to trying to play the signaling game once for each state. If communication has been successful, the choices of sender and receiver that led to success are reinforced by adding $\alpha > 0$ balls of the proper type to the proper urns. If communication has not been successful, then we remove $\beta \geq 0$ balls of the proper types from the proper urns, unless that would remove all balls of that type. This is positive and negative reinforcement, as described above.

On top of that, we will also the effect of *lateral inhibition*. Whenever state t has been successfully communicated with message f, we subtract $\gamma \geq 0$ balls from urn $\mathcal{U}_f$ for each state other than t, as well as the same number of balls from $\mathcal{U}_t$ for each message other than f, unless, again, such a subtraction would remove all balls of a given color from an urn. We call γ the *suppressor*. In the context of a signaling game, we may think of a positive suppressor as implementing biases for synonymy and ambiguity avoidance.

The overall effect of a positive suppressor can be easily foretold: it greatly speeds up the learning process and it will moreover forcefully eradicate sub-optimal options and lead the system more speedily towards pure strategies. We will see presently that a small positive suppressor is also what drives the parallel between BIOT and our models of reinforcement learning for complex cases. Our suggestion therefore is: reinforcement learning with lateral inhibition is a good interpretation of bidirectional optimality from a diachronic point of view. The next section builds up towards this conclusion by a more detailed revision of our simulation results.

3.3 Results

Our simulations show that with only positive reinforcement, some but not all of our critical examples converge to the bidirectionally optimal solutions in most trials, but it may take a very long time to reach convergence. If we include additional negative reinforcement, we get a rather strong but undesirable effect: positive mali drive the learners away from signaling systems, towards babbling and pooling equilibria, and thereby also away from the solutions selected by bidirectional optimality. The key element that assures convergence to the solution selected by bidirectional optimality is lateral inhibition: for sufficiently large λ and a mild bonus, even small suppressor values ensure perfect convergence in remarkably short time.

Effect of Parameter Choices. To give an impression of the learning dynamics under different parameter sets, consider first the case of only positive reinforcement for the OT-system in Figure 1. Figure 5 shows the development of the sender's strategy over 2500 learning steps when $\lambda = 3$, $\alpha = 2$, and $\beta = \gamma = 0$.[8] This trial converges to a partial pooling strategy where f_1 is the dominant choice in states t_1 and t_2. This is not necessarily representative of the dynamics for these parameters, because often the system also falls into total pooling with no communication. Due to the symmetry of the game, the receiver's interpretation behavior evolves in obvious correspondence.

[8] In this and in the following examples, we look at the result of the learning dynamics after a fixed number of learning steps. This is reasonable because we are interested in the most likely outcome on average on a mid-term scale. Reinforcement learning need not converge to pure strategies at all, and unless negative reinforcement or lateral inhibition is allowed to remove all balls from an urn (which we did not allow) there is always a non-zero chance that the system moves away from a pure strategy, no matter how close it got.

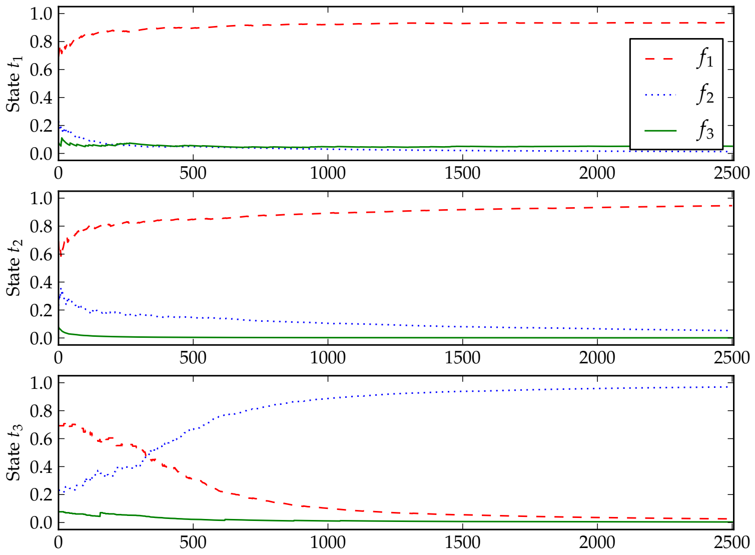


Figure 5: Development of the sender's strategy over 2500 learning steps for the system in Figure 1 with parameters $\lambda = 3$, $\alpha = 2$, and $\beta = \gamma = 0$. Plotted is the development of probabilities with which the sender sends messages in a given state.

Still, this contrasts, with the example shown in Figure 6 which gives a prototypical result for parameter set $\lambda = 3$, $\alpha = 2$, $\beta = 1$, and $\gamma = 0$, where we have positive and negative reinforcement. The trial converges to a partial babbling strategy where the sender sends arbitrary messages in states t_2 and t_3. Occasionally, the dynamics try to break out of the symmetry in state t_2 but negative reinforcement abruptly suppresses these explorations. This behavior is largely stereotypical for negative reinforcement.

As an example of successful learning, Figure 7 gives a prototypical single trial of learning over 800 steps for parameter values $\lambda = 10$, $\alpha = 5$ and $\beta = 0$ and $\gamma = 2$. Initially, the sender's most likely choice is f_1 for all states, but after approximately 500 learning steps this is overtaken by f_2 as the most likely emitted signal for state t_2. Shortly afterwards, after around 600 learning steps, message f_3 very rapidly becomes the most likely emitted signal in state t_3. With lateral inhibition, the system quickly converges to the pure strategies selected by BIOT.

Aggregate Behavior. These are just examples of single learning trials that are not necessarily representative of the aggregate behavior of many runs. We are of course interested in how well the learning dynamics converges to bidirectional optimality for certain parameter configurations *on average*. As the BIOT-algorithm selects sets of meaning-form pairs at each step, which we may take to represent partial strategies,

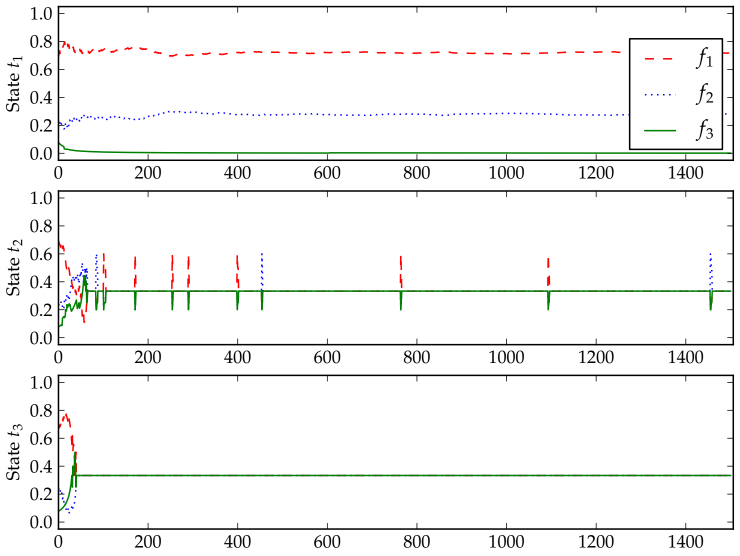


Figure 6: Development of the sender's strategy over 1500 learning steps for the system in Figure 1 with parameters $\lambda = 3$, $\alpha = 2, \beta = 1$, and $\gamma = 0$.

the question is how to measure the similarity between a given total behavioral strategy (the outcome of reinforcement learning) and a partial behavioral strategy (the optimal meaning-form pairs). To determine which learning trials were successful, we looked at all probability distributions given by the partial strategy, i.e, the behavioral strategies for all the choice points that the partial strategy covers. We then computed the *Hellinger similarity*[9] between all of these probability distributions with the ones given by the total strategy for all corresponding choice points. If the Hellinger similarity was at least .8 for all comparisons (an arbitrarily chosen threshold), we considered the learned total strategy to be sufficiently close to the BIOT solution.

Average success rates of this kind are given in Tables 1 through 3. We considered different parameter sets that were most favorable for learning with (i) only positive reinforcement, (ii) positive and negative reinforcement and (iii) positive reinforcement with lateral inhibition. The tables give the numbers of sender and receiver strategies, out of 100 trials, that successfully converged to each solution newly selected as optimal at step n of the BIOT-algorithm, as well as convergence to O. Results are listed for the problematic examples P-Ex 1 and P-Ex 2 from Figures 3 and 4, as well as OT-systems for M-implicatures with 2, respectively 3 meanings and forms (called Horn 2

[9] Hellinger distance is a standard measure of distance between probability distributions P and Q over the same set X, defined as: $H(P, Q) = \sqrt{1 - \sum_{x \in X} \sqrt{P(x) \times Q(x)}}$. Hellinger similarity is defined as $1 - H(P, Q)$ and ranges between 0 and 1, the latter value for identical distributions.

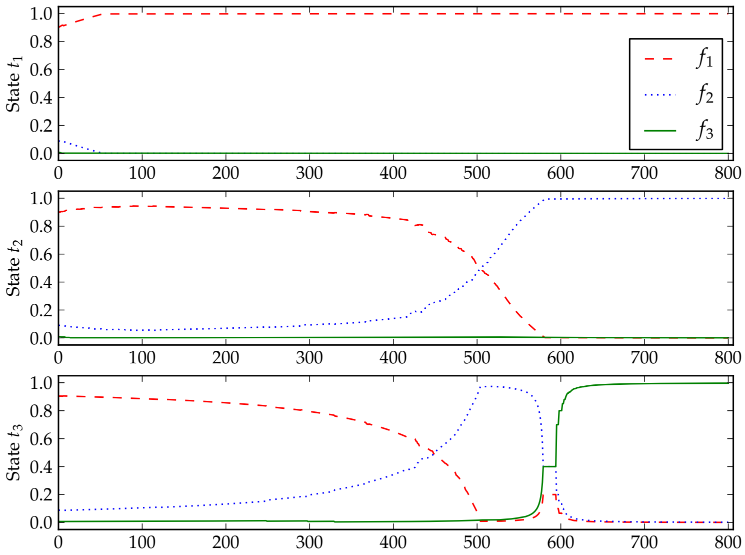


Figure 7: Development of the sender's strategy over 800 learning steps for the system in Figure 1 with parameters $\lambda = 10$, $\alpha = 5$, $\beta = 0$ and $\gamma = 2$.

and Horn 3 here, the latter depicted in Figure 1, the former derived by removing t_3 and f_3 from the latter), and finally, Sc-Imp, the simple OT-system for scalar implicature without QUANTITY-constraint given in example (3).

Learning with only positive reinforcement is slow, occasionally converging to bidirectional optimality, but often meandering along pooling or babbling for a very long time.[10] More specifically, for Horn 2 slightly more trials (ca. 35%) converged to the so-called Anti-Horn strategy (the signaling system that is not selected by BIOT) than to the BIOT-solution (ca. 19%). For Horn 3, the system hardly ever converged to the bidirectionally optimal solution, but occasionally ended up in other signaling systems.

If we look at learning with positive and negative reinforcement, the situation is fairly clear. Negative reinforcement is bad for the evolution of successful signaling, and most trials quickly fall into pooling or babbling. In principle, there is a positive probability of escaping from these —this probability converges to zero as trials proceed along the most likely path of development—, but we failed to observe such "breaking-free" in our simulations. Convergence to BIOT solutions is rare.

Learning with positive reinforcement and lateral inhibition does much better. Still,

[10] Our results are not necessarily indicative of the asymptotic long-term behavior of the dynamics. For example, Argiento et al. (2009)'s analytical result tells us that eventually the strategies in the Horn 2 case should converge to a signaling system. But as the learning rate slows down over time, this long-term effect may never actually show in any trials of our simulations and is also only of marginal practical relevance for reinforcement as model of actual learning (cf. Roth and Erev, 1995).

	Optimality Level							
OT-system	Level 1		Level 2		Level 3		O	
P-Ex 1	100	58	57	0	-	-	55	0
P-Ex 2	48	10	90	50	13	16	13	16
Horn 2	30	19	18	25	-	-	18	19
Horn 3	26	3	6	16	6	10	0	0
Sc-Imp	33	38	-	-	-	-	33	38

Table 1: Rates of convergence to solutions selected by bidirectional optimality (100 trials, 20000 steps, $\lambda = 3$, $\alpha = 3$ and $\beta = \gamma = 0$). The first number gives the percentage of matching sender strategies, the second that of the receiver.

	Optimality Level							
OT-system	Level 1		Level 2		Level 3		O	
P-Ex 1	100	93	90	93	-	-	90	88
P-Ex 2	15	86	14	14	0	0	0	0
Horn 2	41	99	3	3	-	-	3	3
Horn 3	15	87	10	0	0	0	0	0
Sc-Imp	61	0	-	-	-	-	61	0

Table 2: Rates of convergence to solutions selected by bidirectional optimality (100 trials, 4000 steps, $\lambda = 3$, $\alpha = 2$ and $\beta = 1$, $\gamma = 0$).

convergence is not always perfect. In the case of Horn 3, for instance, we find convergence to a different signaling system, and very infrequently also to pooling with only f_1 and t_1. The most likely outcome for this condition also matches the bidirectionally optimal solution in the cases P-Ex 1 and P-Ex 2 that were bothering rationalistic IBR. Despite of all this, the parallel between BIOT and this form of learning in games is also imperfect. It is obvious that this approach is susceptible to frequencies, and so the same kind of "distributional quantity reasoning" that came in the way of a rationalistic grounding of BIOT, also comes in the way here. Under the given parameter choices, learning selects the only possible signaling system for this case in ca. 75% of the trials.

Our conclusion here can only be tentative, as it is based on just a small set of examples. Still, it seems that bidirectional optimality can be characterized rather well as the most likely path of reinforcement learning with lateral inhibition, with the exception of quantity reasoning, which would have to be implemented in BIOT as before by a constraint like in Equation (4) (together with its speaker-side equivalent).

4 Conclusion

Although bidirectional optimization bears an obvious close likeness to game-theoretic models of communication, the devil is in the details. Game-theoretic approaches are often more specific, requiring and manipulating quantitative input where (standard)

	Optimality Level							
OT-system	Level 1		Level 2		Level 3		O	
P-Ex 1	100	100	74	64	-	-	74	64
P-Ex 2	100	100	92	88	88	88	88	88
Horn 2	100	100	100	100	-	-	100	100
Horn 3	100	100	93	92	91	91	91	91
Sc-Imp	0	11	-	-	-	-	0	11

Table 3: Rates of convergence to solutions selected by bidirectional optimality (100 trials, 2000 steps, $\lambda = 10$, $\alpha = 5$ and $\beta = 0$ $\gamma = 2$).

BIOT handles qualitative inputs. Therefore, the main lesson that can be drawn from the mere attempt of comparing BIOT to signaling games is that the latter require explicit commitment to many details that BIOT remains silent about, such as the goals of communication, beliefs, behavioral dispositions etc.

Since there is no unanimity in the literature as to whether BIOT captures the outcome of online processing or of long-term linguistic change, we have tried to recast BIOT both as a rationalistic process of back-and-forth pragmatic reasoning about conversational strategies, and as a gradual process of reinforcement learning. Both comparisons gave a close, but never a perfect fit. Most strikingly, the major difference between BIOT on the one hand and IBR and reinforcement learning on the other is that the latter are sensitive to gaps in the generator, so that "quantity reasoning" is a direct concomitant. In contrast, BIOT requires appropriate constraints to take care of it. As far as we can see, there is neither interpretation of BIOT is clearly superior to the other and therefore there is also no compelling *direct* argument to be derived from the present comparison as to whether the synchronic or the diachronic interpretation of BIOT is more reasonable.

Still, there are nonetheless valuable conceptual lessons to be learned from the proposed comparisons that might indirectly favor either synchronic or diachronic interpretations of BIOT. From a game-theoretic perspective, what is most perplexing about BIOT is its (i) its *fast-and-frugal* association mechanism (sticking to optimal associations once and for all), and (ii) its *lateral blocking behavior*. Our comparisons with IBR and reinforcement learning offer different explanations especially for the latter. Within IBR reasoning lateral blocking occurs when an unexpected message is interpreted *not* as expressing a meaning which might have been expressed by a non-surprising message: a kind of so-called forward induction reasoning. Within reinforcement learning lateral inhibition shows as a bias against synonymy and ambiguity. Either mechanism may be more or less plausible for certain applications of BIOT, but not for others.

Finally, it should be mentioned that IBR and reinforcement learning are only two possible game-theoretic interpretations of BIOT. There might be others, although we presently consider these two the most obvious and plausible. This being said, it is also clear that it is relatively easy to find mechanisms that mimic only strong bidirectional optimality. We were, however, interested in the full generality, and therefore even considered contrived example cases that might not be relevant for linguistic applications.

A Proof of Theorem 2.2

The fact that there are no ties entails that best responses are always unique, with the exception to the best responses to surprise messages. This follows directly from the following lemma:

Lemma A.1.

$$\sum_{t'} \rho(t'|f) u_s(t, f_1, t') = \sum_{t'} \rho(t'|f) u_s(t, f_2, t') \Rightarrow f_1 = f_2$$
$$\sum_{t'} \sigma(t'|f) u_r(t', f, t_1) = \sum_{t'} \sigma(t'|f) u_r(t', f, t_2) \Rightarrow t_1 = t_2$$

Proof. According to the implementation of nominal costs via the definition in Equation (1), the equation in the first line can be rewritten as

$$\langle \rho(t|f), -c(t, f_1)\rangle = \langle \rho(t|f), -c(t, f_2)\rangle\,,$$

which entails that $c(t, f_1) = c(t, f_2)$. Due to NoTies, it follows that $f_1 = f_2$.

The equation in the second line can be rewritten as

$$\langle \sigma(t_1|f), -c(t_1, f)\rangle = \langle \sigma(t_2|f), -c(t_2, f)\rangle\,.$$

This entails that $c(t_1, f) = c(t_2, f)$ and thus $t_1 = t_2$. □

Since the lexicographic ordering over utilities is well-founded for a finite domain, there is always a unique message that maximizes the expected utility for a sender of type t against a behavioral receiver strategy ρ. Likewise, there is always a unique choice the maximizes the expected utility of a receiver upon observing a message f against a behavioral sender strategy σ, provided f is not a surprise message under σ. As a corollary, it follows that $\sigma_n(f|t) \in \{0, 1\}$ for all n.

Now the definition of the IBR sequence (Definition 2.1) can be rewritten as

$$\rho_0(t|f) \quad = \quad |\mathcal{G}(f)|^{-1}$$

$$\sigma_n(f|t) \quad = \quad \begin{cases} 1 & \text{if} \quad \rho_n(t|f) = 1 \\ 1 & \text{if} \quad \forall f' : \rho_n(t|f') < 1 \wedge \\ & \qquad \rho_n(t|f) = |\mathcal{G}(f)|^{-1} \wedge \\ & \qquad \neg\exists f' \prec_t f : \rho_n(t|f') = |\mathcal{G}(f')|^{-1} \\ 1 & \text{if} \quad \forall f' : \rho_n(t|f') = 0 \wedge \neg\exists f' \prec_t f \\ 0 & \text{else} \end{cases}$$

$$\rho_{n+1}(t|f) \quad = \quad \begin{cases} 1 & \text{if} \quad \sigma_n(f|t) = 1 \wedge \neg\exists t' \prec_f t : \sigma_n(f|t') = 1 \\ |\mathcal{G}(f)|^{-1} & \text{if} \quad \forall t' : \sigma_n(f|t') = 0 \\ 0 & \text{else} \end{cases}$$

Let us consider the clauses in detail. The clause for ρ_0 is trivial. We proceed with ρ_{n+1}. Since the prior is uniform, all types that send message f with probability 1 have the same posterior probability. Hence the best response is the type with this property that induces least costs (which is unique; see Lemma A.1). If f is not sent in any type under σ_n, the best response is the uniform distribution over all possible types.

Note that the function $\lambda f \iota t.\rho_{n+1}(t|f) = 1$, if defined, is a bijection. If $\rho_{n+1}(t|f_1) = \rho_{n+1}(t|f_2) = 1$, then $\sigma_n(f_1|t) = \sigma_n(f_2|t) = 1$, hence $f_1 = f_2$.

We now proceed to the re-definition of σ_n. If t can be induced by some message with probability 1 under ρ_n, a best response to t will pick such a message. Due to the considerations from the previous paragraph, this message is unique.

If t is uninducable and there are surprise messages under ρ_n, the maximal posterior probability that can be induced against ρ_n is $|\mathcal{G}(f)|^{-1}$ for some surprise message f. The best response is thus to pick the least costly suprise message. (The constraint in Equation (4) ensures that this will also be the message that induces the highest posterior probability possible.)

If t is uninducible and there are not surprise message, the best response is to pick the least costly message available.

The following lemma captures the implementation of forward induction in the IBR model, i.e. the interpretation of surprise messages:

Lemma A.2. For all $n \in \mathbb{N}$: $\forall t \sigma_n(f|t) = 0$ iff $\forall t \rho_{n+1}(t|f) \neq 1$.

Proof. Suppose $\forall t \sigma_n(f|t) = 0$. It follows directly from the reformulated definition of ρ_{n+1} (second clause) that $\forall t \rho_{n+1}(t|f) \neq 1$. So the direction left to right holds. Now suppose $\exists t : \sigma_n(f|t) = 1$. Then the set $\{t'|\sigma_n(f|t') = 1\}$ has a minimal element with respect to $\prec_f$ due to well-foundedness. Hence $\exists t' : \sigma_n(f|t') = 1 \wedge \neg\exists t'' \prec t' : \sigma_n(f|t'') = 1$. Hence, according to the definition, $\rho_{n+1}(t|f) = 1$, which is a contradiction. Thus the right-to-left direction holds as well. □

It follows that $\rho_n(t|f) = |\mathcal{G}(f)|^{-1}$ iff $\forall t' : \rho_n(t|f) \neq 1$, which hold iff $f \notin Rg(\mathcal{I}_n)$. Given this, we can characterize the relation $S_n = \{\langle t, f\rangle|\sigma_n(f|t) = 1\}$ as follows:

$$\begin{aligned} S_n \;=\;& \mathcal{I}_n \cup \\ & \{\langle t, f\rangle \in \mathcal{G} - Dom(\mathcal{I}_n) \times Rg(\mathcal{I}_n)|\forall f' \prec_t f : f' \in Rg(\mathcal{I}_n)\} \cup \\ & \{\langle t, f\rangle|t \notin Dom(\mathcal{I}_n) \wedge \mathcal{G}(t) \subseteq Rg(\mathcal{I}_n) \wedge \neg\exists f' \prec_t f\} \end{aligned}$$

It follows directly from the definition that $\mathcal{I}_{n+1} = \{\langle t, f\rangle|S_n(t, f) \wedge \neg\exists t' \prec_f t : S_n(t', f)\}$.

Now we can proceed to give the proof of the main theorem via complete induction over n.

- *Induction base* ($n = 0$): According to the definitions, $\mathcal{I}_0 = \emptyset$, and it follows directly that $\{\langle t, f\rangle\, |\rho_0(t|f) = 1\} = \emptyset$ as well.

- *Induction step* ($n > 0$): Suppose $\mathcal{I}_n(t, f)$. Then there is no $t' \prec_f t : S_n(t', f)$. Let us assume otherwise, i.e., there is some $t' \prec_f t$ with $S_n(t', f)$. By induction hypothesis, $\neg\mathcal{I}_n(t', f)$. Since $f \in Rg(\mathcal{I}_n)$, it follows that $t' \notin Dom(\mathcal{I}_n)$, $\mathcal{G}(t) \subseteq Rg(\mathcal{I}_n)$, and $\neg\exists f' \prec_{t'} f$. Now let m be the smallest number such that $t \notin Dom(\mathcal{I}_m)$ and $t \in Dom(\mathcal{I}_{m+1})$. Clearly $m < n$. Then by induction hypothesis, $t' \notin Dom(\mathcal{I}_m) \rightarrow \exists f' \prec_{t'} f' : f \notin Rg(\mathcal{I}_n)$. As $\mathcal{I}_m \subseteq \mathcal{I}_n$ by induction hypothesis,

$t' \notin Dom(\mathcal{I}_m)$. Hence there is an $f' \prec_{t'} f$, which is a contradiction. So the fact that $\mathcal{I}_n(t,f)$ does in fact entail that there is no $t' \prec_f t : S_n(t',f)$, and hence that $\mathcal{I}_{n+1}(t,f)$.

Now suppose $\mathcal{G}(t,f)$, $t \notin Dom(\mathcal{I}_n)$, $f \notin Rg(\mathcal{I}_n)$, and $\forall f' \prec_t f : f' \in Rg(\mathcal{I}_n)$. As $f \notin Rg(\mathcal{I}_n)$, $\neg\exists t' \prec_f t : \mathcal{I}_n(t',f)$. Also, evidently $\mathcal{G}(t) \nsubseteq Rg(\mathcal{I}_n)$. Hence $\neg\exists t' \prec_f t : S_n(t',f)$ iff $\forall t' \prec_f t : t' \notin Rg(\mathcal{I}_n) \rightarrow \exists f' \prec_t f : f' \notin Rg(\mathcal{I}_n)$.

Finally, suppose $t \notin Dom(\mathcal{I}_n)$, $\mathcal{G}(t) \subseteq Rg(\mathcal{I}_n)$, and $\neg\exists f' \prec_t f$. Then there is a $t' \neq t$ with $\mathcal{I}_n(t',f)$. Let m be the number such that $\neg\mathcal{I}_m(t',f)$ and $\mathcal{I}_{m+1}(t',f)$. Clearly $m < n$. By induction hypothesis, $\forall t'' \prec_f t'(t'' \notin Dom(\mathcal{I}_m) \rightarrow \exists f' \prec_{t''} f : f' \notin Rg(\mathcal{I}_m))$. Suppose $t \prec_f t'$. As $t \notin Dom(\mathcal{I}_n)$ and $m < n$, $t \notin Dom(\mathcal{I}_m)$. Hence $\exists f' \prec_t f$, which is a contradiction. Hence $t' \prec_f t$. Since $\mathcal{I}_n(t',f)$, $S_n(t',f)$, so $\exists t' \prec_f t : S_n(t',f)$.

From these considerations it follows that the induction step holds.

This concludes the proof. □

References

Aloni, M. (2007). Expressing ignorance or indifference. Modal implicatures in bi-directional optimality theory. In ten Cate, B. and Zeevat, H., editors, *Logic, Language and Computation: Papers from the 6th International Tbilisi Symposium*, volume 4363, pages 1–20. Springer Verlag, Berlin.

Argiento, R., Pemantle, R., Skyrms, B., and Volkov, S. (2009). Learning to signal: Analysis of a micro-level reinforcement model. *Stochastic Processes and their Applications*, 119:373–390.

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

Benz, A. and van Rooij, R. (2007). Optimal assertions and what they implicate. *Topoi — an International Review of Philosophy*, 27(1):63–78.

Blutner, R. (1998). Lexical pragmatics. *Journal of Semantics*, 15(2):115–162.

Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17(3):189–216.

Blutner, R. and Zeevat, H., editors (2004). *Optimality Theory and Pragmatics*. Palgrave MacMillan.

Blutner, R. and Zeevat, H. (2008). Optimality-theoretic pragmatics. To appear in: Claudia Maienborn, Klaus von Heusinger and Paul Portner (eds.) *Semantics: An International Handbook of Natural Language Meaning*.

Dekker, P. and van Rooij, R. (2000). Bi-directional optimality theory: An application of game theory. *Journal of Semantics*, 17:217–242.

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. PhD thesis, Universiteit van Amsterdam.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics & Pragmatics*, 4(1):1–82.

Grice, P. H. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics, Vol. 3, Speech Acts*, pages 41–58. Academic Press.

Hendriks, P. and de Hoop, H. (2001). Optimality theoretic semantics. *Linguistics and Philosophy*, 24:1–32.

Hendriks, P., de Hoop, H., Krämer, I., de Swart, H., and Zwarts, J. (2007). Conflicts in interpretation. Unpublished book manuscript, Groningen, Nijmegen, Utrecht.

Horn, L. (1984). Towards a new taxonomy for pragmatic inference: Q-based and R-based implicatures. In Schiffrin, D., editor, *Meaning, Form, and Use in Context*, pages 11–42. Georgetown University Press, Washington.

Jäger, G. (2002). Some notes on the formal properties of bidirectional Optimality Theory. *Journal of Logic, Language and Information*, 11(4):427–451.

Jäger, G. (2008). Game theory in semantics and pragmatics. manuscript, University of Bielefeld.

Levinson, S. C. (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicature*. MIT Press, Cambridge, Massachusetts.

Lewis, D. (1969). *Convention*. Harvard University Press, Cambridge, Mass.

Mühlenbernd, R. (2010). Learning with neighbors: Emergence of convention in a society of learning agents. Unpublished manuscript.

Parikh, P. (1987). *Language and Strategic Inference*. PhD thesis, Stanford University. unpublished.

Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report TR-2, Rutgers University Cognitive Science Center, New Brunswick, NJ.

Roth, A. E. and Erev, I. (1995). Learning in extensive form games: Experimental data and simple dynamical models in the intermediate term. *Games and Economic Behavior*, 8(1):164–212.

Selten, R. (1998). Features of experimentally observed bounded rationality. *European Economic Review*, 42(3-5):413–436.

Skyrms, B. (2010). *Signals*. Oxford University Press.

Stalnaker, R. (2006). Saying and meaning, cheap talk and credibility. In Benz, A., Jäger, G., and van Rooij, R., editors, *Game Theory and Pragmatics*, pages 83–100. Palgrave MacMillan, Hampshire.