Vagueness, Signaling & Bounded Rationality

Michael Franke[1], Gerhard Jäger[1], and Robert van Rooij[2]⋆⋆

[1] University of Tübingen
Tübingen, Germany
[2] Universiteit van Amsterdam & ILLC
Amsterdam, The Netherlands

Abstract. Vagueness is a pervasive feature of natural languages that is challenging semantic theories and theories of language evolution alike. We focus here on the latter, addressing the challenge of how to account for the emergence of vague meanings in signaling game models of language evolution. We suggest that vagueness is a natural property of meaning that evolves when *boundedly rational* agents repeatedly engage in cooperative signaling.

Keywords: vagueness, signaling games, language evolution, bounded rationality, fictitious play, categorization, quantal response equilibrium

1 Introduction

Much of what is said in natural language is vague, and members of almost any lexical category can be vague. The question that naturally arises is *why* vagueness is so ubiquitous in natural languages. This paper tries to give an answer to this question by investigating under which circumstances evolutionary processes that are traditionally consulted to explain the emergence of linguistic meaning give rise to "vague meanings" as opposed to "crisp ones". Before all, let us first make clear what we mean when we speak of vagueness in natural language meaning.

Traditional truth-conditional semantics assumes a principle of *bivalence*: either a given individual has a given property, or it does not. But vague predicates challenge this appealing picture, in that so-called *borderline cases* seemingly necessitate a third option. For example, even if we are competent speakers of English who know that John's height is precisely 180 cm, we may not know whether the sentence "John is tall" is therefore true or false. It may either be that this sentence has no unique, objective bivalent truth value, or it may be that it does but that we do not know it, and perhaps cannot know it for principled reasons. Although there is no consensus in the literature about the metaphysical status of borderline cases, it is nonetheless widely assumed that the existence of borderline cases, while possibly being a necessary condition for vagueness, is not a sufficient one (c.f. [33,19]): for a predicate to be truly vague it seems that not only should it have borderline cases, but also there should be borderline cases of borderline cases, and so on. In other words, genuine vagueness is constituted not by *first-order vagueness* alone but by *higher-order vagueness* with completely blurred

⋆⋆ Author names appear in alphabetical order.

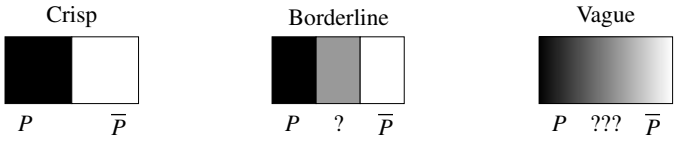

Fig. 1: Crisp and vague denotations, schematically

boundaries. In rough terms, if P is a one-placed predicate and if $\overline{P}$ is its negation, then we would say that these predicates have *crisp denotations* if they split a domain of individuals like shown on the left in Figure 1; existence of a borderline case would look somewhat like in the middle; higher-order vagueness would amount to a gradual blending of categories as depicted on the right.

This paper would like to suggest a *naturalistic explanation* of why and when natural language meanings have vague meanings of this kind. Towards such an explanation, we look at signaling games that were first defined by David Lewis [21] and that have since found good use in modeling the evolution of language (c.f. [29,13,34]). In particular, signaling models are usually employed to demonstrate how linguistic meaning arises from behavioral patterns, which are in turn emerging from repeated interaction in a population of agents. We are interested then in a set of *prima facie* plausible conditions under which a population of signalers converges on a vague code. These conditions would then constitute part of a causal explanation for vagueness: it is *because of* these conditions that (otherwise standard) evolutionary processes lead to vagueness.

Interestingly, a general account of vagueness in terms of the evolution of signaling conventions confronts us with a technical problem. Roughly put, mainstream solution concepts predict that the emerging meaning in adequately defined signaling games will be crisp, because this is *more efficient* and therefore selected for by rationality and evolutionary optimization alike. The main technical question that this paper adresses is therefore: under what reasonable (but conservative) changes to the signaling games framework do vague meanings arise? We focus on models in which signaling agents are *boundedly rational* to some degree, i.e., models which impose some limitations on agents' information processing capabilities. More concretely, we show that vague meanings arise from signaling under two conditions: (i) when agents play rationally but have *limited memory*, and (ii) when agents plays *stochastically rational* due to certain imprecisions in assessing fully optimal play.

The paper is organized as follows. Section 2 introduces the signaling games approach to language evolution and discusses the above mentioned problem that optimization should irrevocably lead to crisp denotations. Section 3 reflects on a number of conceptually plausible explanations for why language might be vague despite this apparent non-optimality. Section 4 then gives a model where the signaling agents play rationally but have only a finite memory of past encounters. Section 5 finally gives a model in which vague meanings arise because the agents' perception of the signaling situation is "noisy" (in a sense to be made clear later on).

2 Signaling games and the Suboptimality of Vagueness

Signaling Games. A signaling game is an extensive game of imperfect information between a sender S and a receiver R. S observes the actual state $t \in T$, but R only knows that state $t \in T$ occurs with probability $\Pr(t) > 0$. S can send a message $m \in M$ to R, after the observation of which R needs to choose an action $a \in A$. The utilities of players $\mathrm{U}_{S,R}\colon T \times M \times A \to \mathbb{R}$ map each outcome, i.e., each triple $\langle t, m, a\rangle$ that constitutes one round of playing the game, to a numeric payoff for both players.

We will look in particular at two kinds of signaling games in this paper. The first one is what we call *signaling games for type matching* and will mainly be used to illustrate basic concepts for better understanding. In type-matching games, players are cooperative, signaling is costless and play is successful if and only if the receiver's action matches the sender's private information. More concretely, these games have $T = A$ and utilities $\mathrm{U}_{S,R}(t, m, t') = 1$ if $t = t'$ and 0 otherwise. We also conveniently assume throughout that $|M| = |T|$ and $\Pr(t) = \Pr(t')$ for all $t, t' \in T$.

When it comes to explaining how higher-order vagueness can arise in natural language meaning, it is not reasonable to call on signaling games for type matching with their crude all-or-nothing utility structure. Rather we will look at what we call here *signaling games for similarity maximizing*, or, for short, *sim-max signaling games*. In these games, success in communication is a matter of degree, indeed, a matter of *how closely* the receiver's action matches the sender's private information. Technically, we again require $T = A$, but we now assume that the state space T comes with a suitable *objective* similarity measure proportional to which utilities are defined. To keep matters simple, we will assume within this paper that $T \subseteq [0; 1]$ is a (usually: finite) subset of the unit interval, and that $\mathrm{U}_{S,R}(t, m, t')$ is identified with the similarity between t and t', which in turn is given by a Gaussian function of their Euclidean distance:

$$\mathrm{sim}(t, t') = \exp\left(\frac{-(t - t')^2}{2\sigma^2}\right). \tag{1}$$

Equilibrium Solutions. Agents' behavior is captured in terms of strategies. A *pure sender strategy* s is a function from T to M that specifies which messages S would send in each state. Similarly, a *pure receiver strategy* r is a function from M to A that specifies how R would react to each message. *Mixed strategies*, denoted by σ and ρ respectively, are functions from choice points to probability distributions over action choices: $\sigma\colon T \to \Delta(M)$ and $\rho\colon M \to \Delta(A)$. The *expected utility* for $i \in \{S, R\}$ of playing mixed strategies σ and ρ against each other is defined as:

$$\mathrm{EU}_i(\sigma, \rho) = \sum_{t \in T} \sum_{m \in M} \sum_{a \in A} \Pr(t) \times \sigma(m \mid t) \times \rho(a \mid m) \times \mathrm{U}_i(t, m, a)\,.$$

A *(mixed) Nash equilibrium* (NE) of a signaling game is a pair of (mixed) strategies $\langle \sigma^*, \rho^*\rangle$ where neither agent would gain from unilateral deviation. Thus, $\langle \sigma^*, \rho^*\rangle$ is an NE iff $\neg\exists\sigma\colon \mathrm{EU}_S(\sigma, \rho^*) > \mathrm{EU}_S(\sigma^*, \rho^*)$ and $\neg\exists\rho\colon \mathrm{EU}_R(\sigma^*, \rho) > \mathrm{EU}_R(\sigma^*, \rho^*)$. An NE is *strict* if any unilateral deviation strictly diminishes the deviating agent's expected utility. Strict NEs are stable resting points of gradual processes of bi-lateral optimization.

Emergent Meaning. Lewis famously argued that strict NEs of signaling games for type matching can be seen as endowing initially meaningless signals with a behaviorally-grounded meaning ([21]). In general, any mixed strategy profile $\langle \sigma, \rho \rangle$ for any given signaling game determines how signals are used by the sender to describe states (via σ), and how the receiver interprets these (via ρ). We therefore define the *descriptive meaning* of an expression m, $F_\sigma(m) \in \Delta(T)$, as the likelihood of states given m and σ, and the *imperative meaning* of m, $F_\rho(m) \in \Delta(A)$, as the probability of actions given m and ρ:

$$F_\sigma(m,t) = \frac{\sigma(m \,|\, t')}{\sum_{t' \in T} \sigma(m \,|\, t')} \qquad\qquad F_\rho(m,a) = \rho(a \,|\, m)\,.$$

Of course, we are particularly interested in the (descriptive and imperative) meanings that the strict NEs of a game give rise to. So, for a concrete example, consider a simple signaling game for type matching with two states $T = \{t_1, t_2\}$ and two messages $M = \{m_1, m_2\}$. This game has only two strict NEs, which are given by the only two bijections from T to M as the sender strategy, and the respective inverses thereof as the receiver strategy. In both NEs descriptive and imperative meaning coincide, as each message comes to denote a unique state, both descriptively and imperatively. In other words, we find exactly two stable "languages" here, characterized by: $F_{\sigma,\rho}(m_{\mathrm{i}}) = t_{\mathrm{k}}$ and $F_{\sigma,\rho}(m_{\mathrm{j}}) = t_{\mathrm{l}}$, where $i, j, k, l \in \{1, 2\}$, $i \neq j$ and $k \neq j$.

Emergent Vagueness? In the previous example, evolved meanings are crisp, not vague: there is no overlap between denotations, no borderline cases, just a clear meaning distinction between messages with disjoint denotations. This is generally the case: it is easy to see that the strict NEs of type matching games, as defined here, *never* give rise to vague meanings with (partially, gradually) overlapping denotations. It is tempting to think that this should be different for sim-max games, where the state space is continuously ordered by objective similarity, and where thus continuous category transitions seem *prima facie* plausible. But this is not so. Sim-max games have been studied by, *inter alia*, [18], [16] and [17] where it is shown that in all strict NEs of these games (a) the imperative meanings of the signals are singular *prototypes*, i.e., designated singular *points* of the type space that best represent a signal's meaning, and (b) the indicative meanings are the *Voronoi tesselations* that are induced by these prototypes.

These results for sim-max games are to a large extent very encouraging because they directly correspond to several findings of cognitive semantics (cf. [8]), but it is nowhere near an account of vagueness. For that we would like to see "blurry tesselations" with gradual prototypicality and gradual category membership. Douven et al. ([4]) essentially consider the same problem when they try to integrate vagueness into the conceptual spaces framework of [8]. They do so by constructing tesselations with thick but precise category boundaries from extended but precise prototype *regions*. Our approach is in a sense more ambitious. Firstly, we would also like to include *gradation* in prototypicality and category membership, so as to capture higher-order vagueness. Secondly, we would also like to *derive* "blurry tesselations" —as opposed to mathematical construction— from properties of linguistic interaction.

The Suboptimality of Vagueness. There is, however, a considerable conceptual obstacle: as the above examples already demonstrated, it holds quite in general that standard models of optimal choice preclude vagueness. In a nutshell, the problem is this (Lipman presents a slightly different, more precise formulation in [22]). Firstly, notice that any pure sender strategy will always give rise to a descriptive meaning with sharp, non-vague boundaries. So, in order to see vagueness emerge, we would minimally need a case where a non-degenerate[3] mixed sender strategy is part of a strict NE. But, secondly, it is also easy to see that no non-degenerate mixed strategy is ever *strictly* better than any of the pure strategies in its support. Phrased the other way around, strict NEs will contain only pure strategies. But that means that a vague language, captured in terms a non-degenerate strategies, is never a stable outcome of evolutionary optimization. As Lipman puts it: vagueness cannot have an advantage over precision and, except in unusual cases, will be strictly worse.

3 Re-Rationalizing Vagueness

Lipman's argument implies that we need to rethink some of the implicit assumptions encoded in the signaling game approach to language evolution if we want to explain how vague meanings can emerge from signaling interaction. Any changes to the model should of course be backed up by some reasonable intuition concerning the origin and, perhaps, the benefit of vagueness in language. Fortunately, such intuitions abound, and we should review some relevant proposals.

To begin with, it is sometimes argued that it is *useful* to have vague predicates like 'tall' in our language, because it allows us to use language in a *flexible* way. Obviously, 'tall' means something different with respect to men than with respect to basketball players. So, 'tall' has a very flexible meaning. This does not show, however, that *vagueness* is useful: vagueness is not the same as context-dependence, and the argument is consistent with 'tall' having a precise meaning in each context.

Some argue that our vague, or *indirect*, use of language might be *strategically optimal* given that some of our messages be diversely interpretable by cooperative and non-cooperative participants. Indeed, using game theoretical ideas one can show (e.g. [27], [15], [1]) that once the preferences of speaker and listener are not completely aligned, we can sometimes *communicate more* with vague, imprecise, or noisy information than with precise information. Interesting as this might be, we find it hard to believe that speaker-hearer conflicts should have quite *such* deep impact on the semantic structure of natural languages, given that communication as such requires crucially a substantial level of cooperation.

Still, occasionally it may indeed be beneficial for both the speaker *and* the hearer to sometimes describe the world at a more coarse-grained level (see for instance [12] and [20]): for the speaker, deciding which precise term to use may be harder than using an imprecise term; for the listener, information which is too specific may require more effort to analyze. Another reason for not always trying to be as precise as possible is that this would give rise to *instability*. As stressed by [30], for instance, in case one

[3] We say that a pure strategy is *degenerate* if it is essentially a pure strategy, i.e., if it puts all probability mass on one pure strategy in its support.

measures the height of a person in all too much detail, this measure might change from day to day, which is not very useful. Though all these observations are valid, none of them make a strong case for vagueness. To economize processing effort, language users could equally well resort to precise but less informative, more general terms whenever conversational relevance allows (and if precision is relevant, processing economy would have to be sacrificed anyway). Similar arguments would also apply to the stability of a precise language.

It is natural to assume that the existence of vagueness in natural language is *unavoidable*. Our powers of discrimination are limited and come with a margin of error, and it is just not always possible to draw sharp borderlines. This idea is modeled in Williamson's [36] epistemic treatment of vagueness, and given a less committed formulation in [32] using Luce's [23] preference theory. This suggests to explain vagueness in terms of a theory of *bounded rationality*: language is vague because its users have *limited information processing capabilities*. In order to fill this general idea with life, we would like to investigate two particular hypotheses. Firstly, we conjecture that vague meanings arise in signaling games if interlocutors have only a finite recollection of previous interactions (Section 4). Secondly, we suggest in Section 5 that vague meanings also show in signaling game models if agents choose actions with a probability proportional to its expected utility. The motivation for this approach is that there might be systematic noise somewhere in the agents' assessment of optimal behavior, be it either in the agents' perception of the game's payoff structure, in the agents' calculation of expected utilities, or yet something we, as modellers, are completely unaware of.

4 Limited Memory Fictitious Play

Fictitious play in normal form games. Humans acquire the meanings of natural language signals (and other conventional signs) by *learning*, i.e., by strategically exploiting past experience when making decisions. A standard model of learning in games is *fictitious play* (see [2]). In its simplest incarnation, two players play the same game against each other repeatedly an unlimited number of times. Each player has a perfect recall of the behavior of the other player in previous encounters, which makes for a loose parallel of this dynamics with exemplar-based theories of categorization (cf. [28]). The players operate under the assumption that the other player is stationary, i.e., he always plays the same —possibly mixed— strategy. The entire history of the other player's behavior is thus treated as a sample of the same probability distribution over pure strategies. Using Maximum Likelihood Estimation, the decision maker identifies probabilities with relative frequencies and plays a best response to the estimated mixed strategy. Most of the research on this learning dynamics has focused on normal form games, where strict NEs are provably absorbing states. This means that two players who played according to a certain strict NE will continue to do so indefinitely. Also, any pure-strategy steady state must be an NE. Furthermore, if the relative frequencies of the strategies played by the agents converge, they will converge to some (possibly mixed strategy) NE. For large classes of games (including 2x2 games, zero sum games, and games of common interest) it is actually guaranteed that fictitious play converges (see [6], Chapter 2, for an overview of the theory of fictitious play and further references).

Limited memory. This result rests on the unrealistic assumption that the players have an unlimited memory and an unlimited amount of time to learn the game. In a cognitively more realistic setting, players only recall the last n rounds of the game, for some finite number n. We call the ensuing dynamics the *limited memory fictitous play* (LMF) dynamics. For the extreme case of $n = 1$, LMF dynamics coincides with so-called Cournot dynamics in strategic games (see Chapter 1 of [6]).

In strategic games LMF dynamics preserves some of the attractive features of fictitious play. In particular, strict NEs are absorbing states here as well. Also, if LMF converges to a pure strategy profile, this is an NE. However, if a game has more than one NE, the memories of the players need not converge at all. To see why, assume that $n = 1$ and the sequence starts with the two players playing different strict NEs. Then they will continue to alternate between the equilibria and never converge to the same NE. Neither is it guaranteed that the relative frequencies of the entire history converge to an NE, even if they do converge. To illustrate this with a trivial example, consider the following coordination game:

	L	R
T	1;1	0;0
B	0;0	2;2

If the dynamics starts with the profile (B, L), the players will alternate between this profile and (T, R) indefinitely. The empirical frequencies will thus converge towards $(\frac{1}{2}, \frac{1}{2})$, which is not an NE of this game.

LMF in Signaling games. There are various ways how to generalize LMF dynamics to signaling games. Observing a single run of an extensive game does not give information about the behavioral strategies of the players in information sets off the path that has actually been played. In some versions of extensive form fictitious play, it is assumed that players also have access to the information how the other player would have played in such unrealized information sets (c.f. [11]). Here we pursue the other option: each player only memorizes observed game histories. We furthermore assume that receivers know the prior probability distribution over types and are Bayesian reasoners. Finally, we assume that both players use the *principle of insufficient reason* and use a uniform probability distribution over possible actions for those information sets that do not occur in memory.

To make this formally precise, let $\bar{s} \in (T \times M)^n$ be a sequence of type-signal pairs of length n. This models the content of the receiver's memory about the sender's past action. Likewise $\bar{r} \in (M \times T)^n$ models the sender's memory about the receiver's past action. We write $\bar{s}(k)$ and $\bar{r}(k)$ for the k^{th} memory entry in $\bar{s}$ or $\bar{r}$. These memories define mixed strategies as follows:

$$\sigma(m \mid t) = \begin{cases} \frac{|\{k \mid \bar{s}(k) = \langle t, m \rangle\}|}{|\{k \mid \exists m' : \bar{s}(k) = \langle t, m' \rangle\}|} & \text{if divisor } \neq 0 \\ \frac{1}{|M|} & \text{otherwise} \end{cases}$$

$$\rho(t \mid m) = \begin{cases} \frac{|\{k \mid \bar{r}(k) = \langle m, t \rangle\}|}{|\{k \mid \exists t' : \bar{r}(k) = \langle m, t' \rangle\}|} & \text{if divisor } \neq 0 \\ \frac{1}{|T|} & \text{otherwise.} \end{cases}$$

When computing the posterior probability $\mu(t \mid m)$ of type t given signal m, the receiver uses Bayes' rule and the principle of insufficient reason. (As before, $\Pr(\cdot)$ is the prior probability distribution over types.)

$$\mu(t \mid m) = \begin{cases} \frac{\sigma(m \mid t) \Pr(t)}{\sum_{t'} \sigma(m \mid t') \Pr(t')} & \text{if divisor } \neq 0 \\ \frac{1}{|T|} & \text{otherwise.} \end{cases}$$

Best response computation is standard:

$$\mathrm{BR}_S(t; \rho) = \arg\max_m \sum_{t' \in T} \rho(t' \mid m) \times \mathrm{U}_S(t, m, t')\,,$$

$$\mathrm{BR}_R(m; \mu) = \arg\max_t \sum_{t' \in T} \mu(t' \mid m) \times \mathrm{U}_R(t', m, t)\,.$$

Characterization & Results. How does the LMF dynamic look like in signaling games for type matching? Consider the basic 2-state, 2-message game, with its two strict NEs. It turns out that these equilibria are absorbing states under fictitious play with unlimited memory. However, this does not hold any longer if memory is limited and the game has more than two types. For illustration, assume a signaling game for type matching with three types, t_1, t_2 and t_3, and three forms, m_1, m_2 and m_3. Suppose furthermore that at a certain point in the learning process, both players have consistently played according to the same equilibrium for the last n rounds — say, the one where t_i is associated with m_i for $i \in \{1, 2, 3\}$. With a positive probability, nature will choose t_1 n times in a row then, which will lead to a state where $\bar{s}$ contains only copies of $\langle t_1, m_1 \rangle$, and $\bar{r}$ only copies of $\langle m_1, t_1 \rangle$. If nature then chooses t_2, both m_2 and m_3 will have the same expected utility for the sender, so she may as well opt for m_3. Likewise, t_2 and t_3 have the same expected utility for the receiver as reaction to m_3, so he will choose t_2 with probabilty $1/2$. If this happens, the future course of the game dynamics will gravitate towards the equilibrium where t_2 is associated with m_3, and t_3 with m_2.

Such transitions can occur between any two signaling systems with positive probability. Thus the relative frequencies of actions, if averaged over the entire history, will converge towards the average of all signaling systems, which corresponds to the pooling equilibrium. If the size of the memory is large in comparison to the number of types, this may hardly seem relevant because the agents will spend most of the time in some signaling system, even though they may switch this system occasionally. However, if the number of types is large in comparison to memory size, LMF dynamics will never lead towards the vicinity of a strict equilibrium, even if such equilibria exist.

This observation is not really surprising. In an NE of a signaling game for type matching, the best response to one type does not carry information about the best response to another type (beyond the fact that these best responses must be different). If the agents only have information about a subset of types available in their memory, there is no way how to extrapolate from this information to unseen types. However, if the type space has a topological structure, as in the class of sim-max games, it is actually possible to extrapolate from seen to unseen types to some degree. Similar types lead to similar payoffs. Therefore the information about a certain type is not entirely lost if it intermittently drops out of memory. Likewise, LMF players are able to make informed guesses about

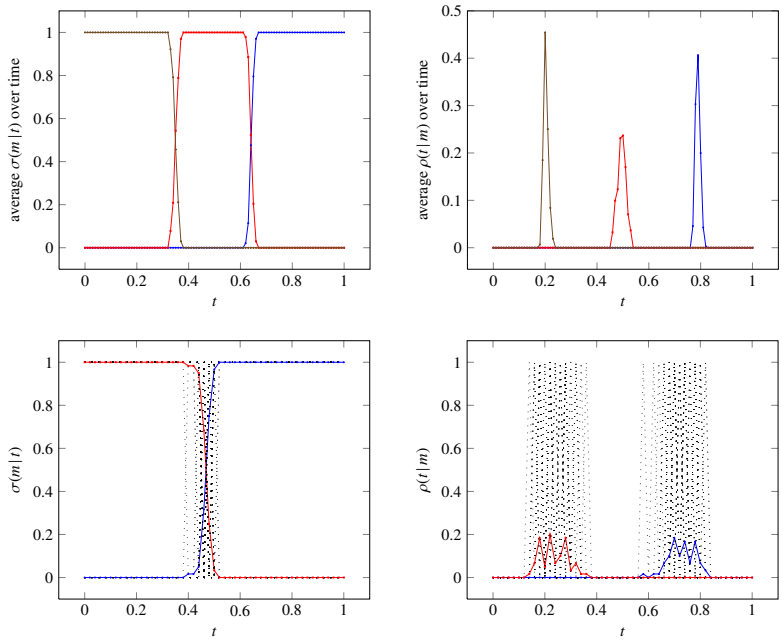

Fig. 2: Results of LMF dynamics

the nature of types that have never been observed before. Consequently, LMF dynamics performs far better in these games. It does not converge towards a strict equilibrium, but somewhere into the proximity of one, thus ensuring a high degree of efficiency.

The top of Figure 2 depicts the outcome of a first simulation of the LMF dynamics. The type space consisted of 500 types that were distributed evenly over the unit interval, and we assumed three signals. The simulation assumed $\sigma = 0.1$ in Equation (1) and a memory size $n = 200$. The graphs show the relative frequencies between the 10,000$^{\text{th}}$ and the 20,000$^{\text{th}}$ iterations of the game, starting from an initial state where the memories of the agents contain random associations. The sender strategies, shown on the top left of Figure 2, induce a partition of the type space into three categories, one for each message. In the long run, these categories partition the type space into three continuous intervals of about equal size. These intervals are largely stable, but the boundaries shift back and forth somewhat over time. Averaging over a longer period thus leads to categories with blurred boundaries. The prototypes of the categories, i.e., the receiver's interpretation of the three signals as shown on the top right, fall into the center of the corresponding category. Again we observe a certain amount of indeterminacy. Over time, the prototypes are distributed according to a bell shaped curve in the center of the corresponding category.

Interpretation. If we look at the properties of the language that emerges under LMF dynamics over a longer course of time, we find that the emerging categories indeed

have non-sharp boundaries, and that they blend seamlessly into one another. On this level of abstraction, the model derives the crucial features of higher-order vagueness that standard signaling models preclude. But the down-side of this model seems to be that although the time-averaged language shows the relevant vagueness properties, the beliefs and the rational behavior of agents *at each time step* do not. For instance, at a fixed time step the sender would use message m_i for all states in the half-open interval $[0; x)$ and another message m_j for any state $> x$. The point-value x would be an infinitesimal borderline case.[4]

The problem that vagueness only shows over time, so to speak, can be overcome by looking at a population of more than two agents playing LMF dynamics. If each of several agents has her own private limited memory, then differences between private memories blur meaning boundaries if we look at the averaged population behavior even at a single moment of time. The bottom half of Figure 2 shows a population average of 60 LMF agents with a memory of 25 past interaction for a sim-max game with 51 states and 2 messages, obtained after 2500 interaction steps. The solid lines are the population averages and the dotted lines give the strategies of each individual agent. This is then an example of a language whose terms are vague because their meaning is bootstrapped from a number of slightly different individual strategies.

How good an explanation of vagueness is this? Conceptually speaking, it is certainly plausible that properties of a language at large emerge from the (limited) power of its users. Moreover, this account is similar in essence to Lewis' approach to vagueness [21], as well as to super- and subvaluation accounts ([5,14]). Here and there vagueness is explained as the result of adding multiple slightly different precise language uses together. But, still, the question is whether memory limitations alone are sufficient to provide a reasonable explanation for vagueness. We do not think so. This is because, although LMF dynamics may explain why language *as such* is vague, each agent's still commands a crisp language at each moment in time. This would leave entirely unexplained the hesitance and insecurity of natural language users in dealing with borderline cases.

The residual problem here is that the notion of a rational best response to a belief —be it obtained from finite observations or otherwise— will *always* yield sharp boundaries and point-level borderline cases. To overcome this problem, and to derive vague meanings also in the beliefs and behavior of individual agents we really need to scrutinize the notion of a rational best response in more detail. The following section consequently discusses a model in which agents play *stochastic best responses*.

5 Quantal Response Equilibria

Stochastic choice rules have been studied extensively in psychology, but have recently also been integrated into models of (boundedly-rational) decision making from economics. We start by providing a sketch of the relevant background on individual choice from psychology, then take this to interactive choices, and finally report on simulation data showing how equilibria of stochastic choices give rise to vague meanings.

[4] This is not entirely correct parlor, since the simulation only approximates a continuous state set. But the point should be clear nonetheless.

Individual Stochastic Choice. When faced with a choice among several alternatives, people often not only indecisive, but even inconsistent in that they make different choices under seemingly identical conditions, contrary to the predictions of classical rational choice theory. Choice behavior is therefore often modeled as a probabilistic process. The idea is that people *do* make rational and consistent decisions, but that there is something in their behavior which we cannot name and which our models do not encode explicitly. That unknown something, however, might be rather systematic, and we can therefore often describe empirically observed choice data as rational given a specific probabilistic source of error.

The general idea is this.[5] Suppose we force subjects to repeatedly make *binary choices* between options a and b under identical conditions. This could either be a classical behavioral choice, such as whether to buy a or b, or it could be a *perceptual choice*, such as which of a or b is louder, heavier etc. Even if we knew all physical properties of a and b, it would be ludicrous to assume that we know every single factor that guides a subject's current choice. In order to account for these uncertain factors, probabilistic choice models assume that subjects do not actually treat a and b as they are represented in the model but rather as a' and b' which are systematically related to a and b but not necessarily the same. For example, if $a, b \in [0, 1]$, we would assume that a is treated as if it was $a + \epsilon$ where ϵ is a "tremble" that is drawn from some probability distribution. It is natural to assume that small trembles are likely and large trembles unlikely. In that case, if a and b are nearly identical, confusion is rather likely, but the more a and b differ, the less likely a mix-up. These assumed trembles can have many causes, also depending on the kind of choice situation we are looking at. Among other things, it could be that we, as modelers, are not fully aware of the choice preferences of our subjects, or it may be that subjects make mistakes in perceiving stimuli for what they are. From the modeller's perspective, experimental choice data can then be deemed, if not fully rational, then at least consistent with a particular distribution of trembles.

More concretely, if we assume that "trembles" with which agents perceive the quality of their choices are drawn from an extreme-value distribution (roughly: small trembles very frequent, large trembles highly unlikely), then choice behavior can be modeled by a so-called *logit probabilistic choice rule* which states that the probability $P(a)$ of selecting a is an exponential function of a's utility $u(a)$ (see [25], [26] and [9] for details):

$$P(a) = \frac{\exp(\lambda u(a))}{\sum_b \exp(\lambda u(b))}\,. \tag{2}$$

Here, $\lambda \geq 0$ captures inversely the influence of the trembles. In other words, λ measures inversely the degree of rationality of the decision maker, where what is rational is defined by the model *without* the trembles. In this sense, $\lambda = 0$ corresponds to a completely irrational agent that picks each action with equal probability regardless of utility. As λ increases to ∞, the probability of non-optimal choices converge to 0, and all optimal choices have equal probability.

[5] We do not mean to suggest that we are faithful to the vast statistical literature on this topic, but we merely wish to motivate our modeling approach in accessible terms. The interested reader is referred to the classics, such as [35] or [24].

Quantal Response Equilibrium. This much concerns a single agent's decision. But stochastic choice models of this kind have more recently also been applied in behavioral game theory to model subjects' *interactive choice behavior*. In that case, systematic imperfection in individual decision making may also systematically alter the structure of equilibria that ensue in strategic situations where all players choose with a globally fixed λ.[6] If in a strategic setting all players use rule (2) with the same value for λ, and all players are correct in assessing the probabilities of each other's behavior, the mixed strategies of the players form a so-called *logit equilibrium*. It can be shown that in games with finitely many strategies, such an equilibrium (also called *quantal response equilibrium* (QRE) in this case) always exists [25,26,10].[7]

Consider as an example a 2-state, 2-message signaling game for type-matching. We represent a mixed sender strategy as a 2×2 matrix P, where p_{ij} gives the relative probability that the sender will send signal m_j if she has type t_i. Likewise, a mixed receiver strategy is represented by a 2×2 matrix Q, with q_{ij} being the probability that the receiver will choose action a_j upon observing signal m_i. For (P, Q) to form a QRE, it must hold that:

$$p_{ij} = \frac{\exp(\lambda q_{ji})}{\sum_k \exp(\lambda q_{ki})} \qquad \text{and} \qquad q_{ij} = \frac{\exp(\lambda p_{ji})}{\sum_k \exp(\lambda p_{ki})}\,.$$

Using these equations and the fact that P and Q are stochastic matrices, it can be shown by elementary calculations that $p_{11} + p_{21} = 1$ and $q_{11} + q_{21} = 1$, and hence that $p_{11} = p_{22}, p_{12} = p_{21}, q_{11} = q_{22}$, and $q_{12} = q_{21}$. From this it follows that $p_{11} = f_\lambda(q_{11})$ and $q_{11} = f_\lambda(p_{11})$, where

$$f_\lambda(x) = \frac{\exp(\lambda x)}{\exp(\lambda x) + \exp(\lambda(1-x))}\,. \tag{3}$$

Now suppose $p_{11} < q_{11}$. f_λ is strictly monotonically increasing. Hence $f_\lambda(p_{11}) = q_{11} < p_{11} < f_\lambda(q_{11})$, and vice versa. These are contradictions. It thus follows that $p_{11} = q_{11}$, i.e. $P = Q$. The entire equilibrium is thus governed by a single value α, where $\alpha = p_{11} = p_{22} = q_{11} = q_{22}$. α is a fixed point of f, i.e., $\alpha = f_\lambda(\alpha)$.

For $\lambda \in [0, 2]$, there is exactly one fixed point, namely $\alpha = 0.5$. This characterizes a *babbling equilibrium* where each message is sent with equal probability by each type, and each action is taken with equal probability regardless of the message received. If $\lambda > 2$, $\alpha = 0.5$ continues to be a fixed point, but two more fixed points emerge, one in the open interval $(0, 0.5)$ and one in $(0.5, 1)$. As λ grows, these fixed points converge towards 0 and 1 respectively. They correspond to two *noisy separating equilibria*. Even though each message is sent with positive probability by each type in such a QRE (and each action is induced by each signal with positive probability), there is a statistical correlation between types, messages and actions. In other words, in these QREs information transmission takes place, even though it is imperfect.

Generalization. Already in this simple example there is no longer a sharp delineation in (descriptive and imperative) meanings of signals, and it turns out that if we attend to the more interesting case of sim-max games, logit equilibria also indeed give rise to

[6] See [31] for a model that dispenses with the homogeneity of λ among players.

[7] As λ goes to infinity, QREs converge to some NE, the limit cases of perfect rationality.

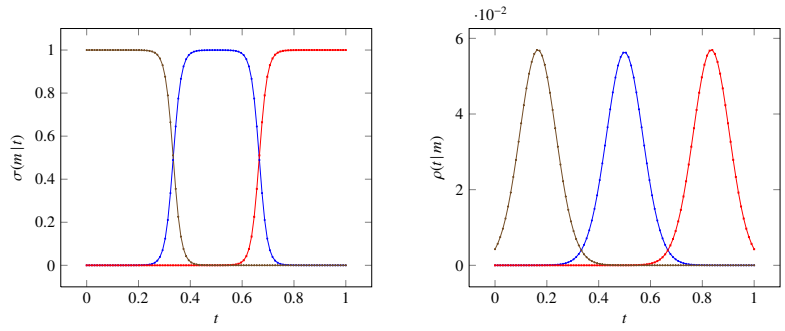

Fig. 3: Separating quantal response equilibrium

continuously blended category boundaries of the relevant kind. We show this by simulation, based on a game with 100 states that are arranged in the unit interval with equal distances. We considered three signals, and chose the value $\sigma = .2$. If λ is small, there is only a (theoretically unappealing) babbling equilibrium in which the sender sends all messages with equal probability in each state. But for values of λ above approximately 4, separating equilibria emerge. Figure 3 shows such an equilibrium for $\lambda = 20$. Here the sender strategy roughly partitions the type space into three categories of about equal size. Crucially, the boundaries between the categories are blurred; category membership smoothly changes from (almost) 1 to (almost) 0 as one moves into a neighboring category. The left half of the figure shows the receiver strategy, i.e., the location of the prototypes. These are not sharply defined points within conceptual space either. Rather, the location of the prototypes can be approximated by a normal distribution with its mean at the center of the corresponding category. In other words, we not only find continuously blended category boundaries in the declarative meaning of signals, but also "graded protoypes" in the imperative meaning.

Interpretation. Vague interpretations of signals emerge with necessity if the perfectly rational choice rule of classical game theory is replace by a cognitively more realistic probabilistic choice rule like the logit choice rule. Unlike for the finite memory model from Section 4, this holds true, also for any momentary belief and behavior of individual agents and even if there are only two agents in a population. The more general reason why this model gives rise to vague meanings is also natural: the sender may only imperfectly observe the state that she wants to communicate, she may make mistakes in determining her best choice, or there may be choice-relevant factors that are not included in the model; similarly for the receiver.

Conceptually, the QRE account of vagueness relates most directly to epistemic accounts of vagueness (e.g. [36]). Since it is natural to assume that players know their opponents' behavior in equilibrium, players should be aware that language is used with a margin of error. Uncertainty about language use is therefore a basic feature of this model. But the QRE model leaves quite some room as to what kind of uncertainty this is. Slack in best responding could come from imprecise observation, but also from con-

textual variability of the preferences of agents. The former would explain quite readily why especially observational predicates are vague, the latter relates this approach to more pragmatic explanations of vagueness (c.f. [7,3]).

Of course, the QRE raises a number of fair questions too. Even if we accept that all natural language expressions are vague, then it is still not necessarily the case that all natural language expressions are vague in the same way: terms like 'red', 'wet' or 'probable' are more readily vague, so to speak, than terms like 'CD-ROM', 'dry' or 'certain'. In further research it would be interesting to relate these properties of meanings to (i) the source and nature of probabilistic error in QRE, and/or to (ii) more nuanced topological properties of the space given by T and the utility function U. Further issues for future research are to extend the two-agent QRE models to more realistic multi-agent models, to combine LMF and QRE, and to take the step from simulation to analytic results where feasible.

References

1. Blume, A., Board, O.: Intentional vagueness (2010), unpublished manuscript, University of Pittsburgh
2. Brown, G.W.: Iterative solutions of games by fictitious play. In: Koopmans, T.C. (ed.) Activity Analysis of Production and Allocation. Wiley, New York (1951)
3. Cobreros, P., Egré, P., Ripley, D., van Rooij, R.: Tolerant, classical, strict. Journal of Philosophical Logic (2011)
4. Douven, I., Decock, L., Dietz, R., Egré, P.: Vagueness: A conceptual spaces approach (2009), unpublished manuscript
5. Fine, K.: Vagueness, truth and logic. Synthese 30(3–4), 265–300 (1975)
6. Fudenberg, D., Levine, D.K.: The Theory of Learning in Games. MIT Press (1998)
7. Gaifman, H.: Vagueness, tolerance and contextual logic. Synthese 174(1), 5–46 (2010)
8. Gärdenfors, P.: Conceptual Spaces: The Geometry of Thought. MIT Press (2000)
9. Goeree, J.K., Holt, C.A.: Stochastic game theory: For playing games, not just for doing theory. Proceedings of the National Academy of Sciences 96(19), 10564–10567 (1999)
10. Goeree, J.K., Holt, C.A., Palfrey, T.R.: Quantal response equilibrium. In: Durlauf, S.N., Blume, L.E. (eds.) The New Palgrave Dictionary of Economics. Palgrave Macmillan, Basingstoke (2008)
11. Hendon, E., Jacobsen, Sloth, B.: Fictitious play in extensive form games. Games and Economic Behavior 15(2), 177–202 (1996)
12. Hobbs, J.: Granularity. In: Proceedings of the International Joint Conference on Artificial Intelligence (1985)
13. Huttegger, S.M.: Evolution and the explanation of meaning. Philosophy of Science 74, 1–27 (2007)
14. Hyde, D.: From heaps and gaps to heaps and gluts. Mind 106, 641–660 (1997)
15. de Jaegher, K.: A game-theoretic rationale for vagueness. Linguistics and Philosophy 26(5), 637–659 (2003)
16. Jäger, G.: The evolution of convex categories. Linguistics and Philosophy 30(5), 551–564 (2007)
17. Jäger, G., Koch-Metzger, L., Riedel, F.: Voronoi languages (2009), to appear in *Games and Economic Behavior*
18. Jäger, G., van Rooij, R.: Language stucture: Psychological and social constraints. Synthese 159(1), 99–130 (2007)

19. Keefe, R., Smith, P. (eds.): Vagueness: A Reader. MIT Press (1997)
20. Krifka, M.: Approximate interpretation of number words: A case for strategic communication. In: Bouma, G., Krämer, I., Zwarts, J. (eds.) Cognitive Foundations of Interpretation, pp. 111–126. KNAW, Amsterdam (2007)
21. Lewis, D.: Convention. A Philosophical Study. Harvard University Press (1969)
22. Lipman, B.L.: Why is language vague? (2009), manuscript, Boston University
23. Luce, D.R.: Semiorders and a theory of utility discrimination. Econometrica 24, 178–191 (1956)
24. Luce, D.R.: Individual Choice Behavior: A Theoretical Analysis. Wiley, New York (1959)
25. McKelvey, R.D., Palfrey, T.R.: Quantal response equilibria for normal form games. Games and Economic Behavior 10(1), 6–38 (1995)
26. McKelvey, R.D., Palfrey, T.R.: Quantal response equilibrium for extensive form games. Experimental Economics 1, 9–41 (1998)
27. Myerson, R.B.: Game Theory: Analysis of Conflict. Harvard University Press (1991)
28. Nosofsky, R.M.: Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General 115(1), 39–57 (1986)
29. Nowak, M.A., Krakauer, D.C.: The evolution of language. PNAS 96, 8028–8033 (1999)
30. Pinkal, M.: Logic and the Lexicon. Kluwer (1995)
31. Rogers, B.W., Palfrey, T.R., Camerer, C.: Heterogeneous quantal response equilibrium and cognitive hierarchies. Journal of Economic Theory 144(4), 1440–1467 (2009)
32. van Rooij, R.: Vagueness and linguistics. In: Ronzitti, G. (ed.) Vagueness: A Guide. Springer (2010)
33. Sainsbury, M.: Is there higher-order vagueness? The Philosophical Quarterly 41(163), 167–182 (1991)
34. Skyrms, B.: Signals. Oxford University Press (2010)
35. Thurstone, L.L.: Psychophysical analysis. American Journal of Psychology 38, 368–389 (1927)
36. Williamson, T.: Vagueness. Routledge (1994)