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 language, but indeed one that is troubling for leading theories in semantics and language evolution. 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.

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

1 Introduction

Vagueness is a pervasive feature of natural language that challenges linguistic theory in manifold ways. For example, according to truth conditional semantics —the most successful and productive linguistic theory of meaning we know so far— the meaning of a declarative sentence is identified with the *conditions*, or *circumstances* under which the sentence is true. The phenomenon of vagueness challenges this view. Even if we know that John's height is 1.80 meters, it is still not clear whether we should count 'John is tall' as being true or as being false.

A traditional way of thinking about vagueness is in terms of the *existence of borderline cases*. John is a borderline case of a tall man, if the sentence 'John is a tall man' is neither (clearly) true nor (clearly) false. The three-valued logic account of vagueness, as well as supervaluation theories without something like Kit Fine's [4] treatment of higher-order vagueness, are based on exactly this idea. Consequently, these theories assume that predicates like 'tall' and 'bald' do give rise to a three-fold partition of objects: the positive ones, the negative ones, and the borderline cases. But it is generally assumed that the existence of borderline cases is inadequate to characterize vagueness: although by assuming a three-fold instead of a two-fold distinction one rightly rejects the existence of a clear border between the positive and the negative cases, one still assumes the existence of an equally unnatural border between, for instance, the positive and the borderline cases. What seems to characterize vagueness, instead, is the fact that the denotation of vague terms lacks sharp boundaries. In the words of Sainsbury [27], they are 'boundaryless': there is no sharp boundary that marks the things which fall under it from the things that do not, and no sharp boundary which marks the things which do definitely fall under it from those which do not definitely, and so on. Instead, all these boundaries are blurred.

** Author names appear in alphabetical order.

Much of what is said in language is vague. Members of almost any lexical category can be vague. This raises the question *why* vagueness is so pervasive in natural languages. Interestingly enough, this is a typical question an *economist* would pose, and she would seek to answer this question in terms of information transmission. It seems obvious that sharing more factual information is always preferred in a cooperative communication setting, meaning that vagueness *cannot* have an advantage over preciseness. The problem, then, is to explain the prevalence of vague terms in natural language. The aim of this paper is to explain this prevalence.

On closer look, the question after the prevalence and origin of vagueness presents itself as a technical problem for our presently best formal models of the evolution of meaning. David Lewis [16] defined the notion of a (cheap talk) signaling game in order to explain how linguistic meaning can arise merely from repeated interaction without assuming any pre-existing semantic code. However, in his paper *Why is language vague?* [17], Barton Lipman presents a convincing case that standard signaling game models cannot give a plausible explanation of vagueness' prevalence. Briefly put, he shows that under natural assumptions, a vague language will always be Pareto-dominated by a non-vague one, provided the communicators are rational. The main technical question that this paper adresses therefore is: under what reasonable (but conservative) changes to the signaling games framework do vague meanings arise?

2 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 chose an action $a \in A$. The utilities of players $\mathrm{U}_{S,R} : T \times M \times A \rightarrow \mathbb{R}$ map each outcome, i.e., each triple $\langle t, m, a\rangle$ that constitutes one round of playing the game, on a numeric payoff for both players. Lewis' [16] showed that although messages may initially be meaningless, repeated interaction of sender and receiver may establish a common code in equilibrium play of the game.

To make this more concrete, fix a *pure sender strategy s* as 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 probability distributions on the set of pure strategies. (Pure strategies can also be regarded as degenerate cases of mixed strategies.) The expected utility for $i \in \{S, R\}$ of playing mixed strategies σ and ρ against each other is defined as:

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

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 $\neg\exists\sigma : \mathrm{EU}_S(\sigma,\rho^*) > \mathrm{EU}_S(\sigma^*,\rho^*)$ and $\neg\exists\rho : \mathrm{EU}_R(\sigma^*,\rho) > \mathrm{EU}_R(\sigma^*,\rho^*)$. An NE is *strict* if any unilateral deviation strictly diminishes the deviating agent's expected utility. Strict NEs correspond to evolutionary stable states: stable resting points of gradual processes of bi-lateral optimization.

Equilibria of a signaling game can explain the emergence of meaning as follows. Suppose for simplicity that the signaling game has only two states $T = \{t_1, t_2\}$, two messages $M = \{m_1, m_2\}$ and two actions $A = T$ that correspond to the states. Assume further that $\mathrm{U}_{S,R}(t, m, t') = 1$ if $t = t'$ and 0 otherwise. We call signaling games where $\mathrm{U}_S = \mathrm{U}_R$ and where receiver actions have to match the states for optimal payoff to sender and receiver *signaling games for type matching*. There are only two strict NEs in this particular game for type matching. In both of these only pure strategies are used. The two NEs are given by the only two bijections from T to M as the sender strategy, and the respective inverse thereof as the receiver strategy.

Generally speaking, a strict NE $\langle \sigma, \rho \rangle$ determines the *descriptive meaning* of an expression m as the posterior probability distribution over states after observing m induced by σ. It also determines the *imperative meaning* of m as the probability distribution over actions that the signal induces given ρ. This is easy for simple cases like the above example. Here, descriptive and imperative meanings coincide, and we may moreover abstract from probabilities: the meaning of a m is the set of all states in which m gets send: $[\![m]\!]^{\sigma} = \{t \in T : \exists s : s(t) = m \wedge \sigma(s) \neq 0\}$. In the present example, the two strict NEs would give rise to two sets of meanings of messages: one in which m_1 denotes $\{t_1\}$ and m_2 denotes $\{t_2\}$, and one in which m_1 denotes $\{t_2\}$ and m_2 denotes $\{t_1\}$.

The meanings that evolve in this game are crisp: there is no overlap between denotations, no borderline cases, just a clear meaning distinction between messages with disjoint denotations. So, when would this approach give rise to a vague meaning? The most obvious idea to try are mixed strategies: we could hypothesize that a stable state gives rise to vague meanings iff it involves mixed strategies whose denotations are partially overlapping, indeed blending continuously into each other. But this is excluded for the simple signaling example above: the only stable states involve pure strategies.

It is tempting to think that this is due to the overly simplistic game we have assumed: after all, many perceptual categories (think: color, pitch, pressure, visual perception of a person's height etc.) involve a large, if not infinite state space that is continuous ordered by some psychophysical measure of similarity. This could be modelled, in some due approximation, by assuming that the state space T is given by the unit interval $[0; 1]$. The set of messages is finite and as before $A = T$. Assume further that payoffs are not all-or-nothing but related to a notion of similarity: for concreteness, assume that $\mathrm{U}_{S,R}(t, m, t')$ is identified with similarity between t and t' which in turn is given by a Gaussian function of their Euclidean distance:

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

Similar games of this variety have been studied by, *inter alia*, [14], [12] and [13] where it is shown that NEs of these games are characterized by strategy profiles where (a) the imperative meanings of the signals are *prototypes*, i.e., designated points of the type space, and (b) the indicative meanings are the *Voronoi tesselations* that are induced by these prototypes. This is an encouraging result because it directly corresponds to several findings of cognitve semantics (cf. [6]). But would this more realistic set-up give rise to stable states that are "blurry Voronoi tessellations" that look less like political maps with country boundaries, and more like a geographic map where hills blend into valleys blend into hills and so on?

The discouraging answer is: no, it wouldn't. This is what Lipman's argument tells us [17]. The crux of the argument is that any non-degenerate mixed strategy is never *strictly* better than any of the pure strategies in its support. Hence we cannot hope to uniquely single out a mixed strategy profile as the unique best NE. Hence, in whatever way we construct the utility function, it will not help to explain the prevalence of vagueness of natural language as meanings that arise uniquely and rationally under continuously overlapping mixed strategies.

More concretely, Lipman's argument takes the following form. Let V be the expected utility of the best sender-receiver strategy pair: $V = \max_{\sigma,\rho} \mathrm{EU}(\sigma, \rho)$. It is easy to see that if there is a pair of strategies $\langle \sigma, \rho \rangle$ such that the maximum V is attained, then every pair of pure strategies $\langle s, r \rangle$, such that s and r are in the support of σ and ρ respectively, is a pure NE in which V is the expected payoff. But this means that vagueness cannot have an advantage over specificity 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 of the obvious and some of the recent 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, which means that it 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.

A valid economic suggestion is based on the idea that our vague, or *indirect*, use of language might be partly explained by our intention that some of our messages be diversely interpretable by cooperative versus non-cooperative participants. Indeed, using game theoretical ideas one can show (e.g. [22], [11], [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, it cannot explain the prevalence of vagueness in cooperative communication.

But occasionally it may be beneficial for both the speaker *and* the hearer to sometimes describe the world at a more coarse-grained level (see for instance [10] and [15]): 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 [24], for instance, in case one 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, we don't feel they explain why so many, if not all, *observational predicates* of our language are vague.

In a more recent paper, Kees van Deemter [3] proposes that many natural language concepts are vague, because vagueness facilitates search. It is argued that due to its vagueness, 'tall' partitions the set of (relevant) individuals into *three* instead of just *two* classes: the tall ones, the not tall (or short) ones, and the ones of average length. If we are now informed that 'x is tall', we only have to check one-third of the cases, instead of half of them. This argument is valid, but it has nothing to do with vagueness: the argument only establishes that more fine-grained classifications are preferred (in this respect) to coarse-grained ones. In particular it strongly suggests that it is better *never* to be vague, and always to be very precise, up to very precise degrees. Thus, we feel that van Deemter's argument favors preciseness rather than vagueness.

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 [29] epistemic treatment of vagueness, and given a less committed formulation in [26] using Luce's [18] preference theory. This suggests to explain vagueness in terms of a theory of *bounded rationality*. In particular, we would like to investigate the following two hypotheses: signaling games can model the emergence of vague meaning if (i) interlocutors face memory constraints, or (ii) agents play stochastic best responses (because there is noise in their perception of the payoff-relevant distinctions). To test these hypotheses, section 4 presents a signaling model in which agents best respond to a belief derived from a limited sequence of the opponent's last n choices, and Section 5 finally presents a model in which agents play a *stochastic best choice* because there is predictable noise either in the games payoff structure, or in the agents' calculation of expected utilities.

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. [23]). 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. There it can be shown that strict NEs are 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 [5], 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 [5]).

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 a NE. However, the memories of the players need not converge at all, as soon as a game has more than one NE. 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 $({}^1\!/_2\,, {}^1\!/_2)$, 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 (see [9] for motivation of this decision and technical exploration of the consequences). 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 unsufficient 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:[3]

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

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

When computing the posterior probability $\mu(t|m)$ 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.)

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

Best response computation is standard:

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

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. In that case, the learner's generalizations are more prone to be influenced by the possible partiality of their observations.

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,

[3] Notice that in signaling games mixed strategies can be equivalently defined as probability distributions on choices for each choice point. That's what we do here.

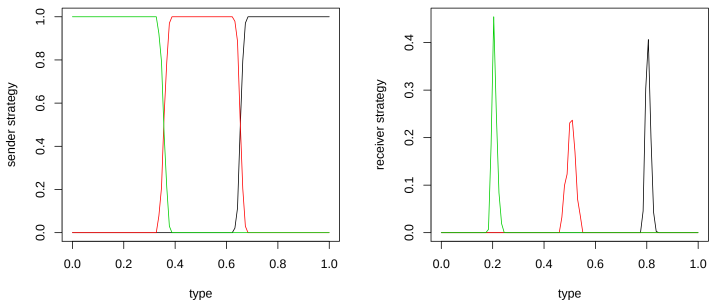

Fig. 1: Long-time average of LMF dynamics

this may hardly seem relevant because the agents will spend most of the time in some signaling system, even though may they 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 games we wish to consider here, 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 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 a strict equilibrium, thus ensuring a high degree of efficiency.

Figure 1 depicts the outcome of a simulation of the LMF dynamics. In simulations continuous space needs to be approximated, and in the present case the type space consisted of 500 types that were spaced evenly over the unit interval, and we assumed three signals. As described in Section 2, the utilities were similarity-sensitive, expressed by a Gaussian function of their Euclidean distance, as in Equation (1). The simulation assumed $\sigma = 0.1$ and a memory size $n = 200$. The graphics depicts the relative frequencies between the 10,000th and the 20,000th iterations of the game, starting from an initial state where the memories of the agents contain random associations. The sender strategies 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 continous 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, 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 some of the crucial features of vagueness that standard signaling models preclude. But is this the right level of analysis?

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 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 it therefore seems that we 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 in psychology, but have recently been integrated into models of (boundedly-rational) decision making from economics. We start by providing a sketch of the relevant background, then discuss the notion of a quantal response equilibrium, and finally report on simulation data showing how equilibria of stochastic choices give rise to vague meanings.

Background: From Stochastic Choice & Categorization. Standard theories of choice assume that strict preference can be modeled by a weak order: an order that is irreflexive, transitive, and negatively transitive.[5] But when faced with a choice among several alternatives, people often do not know what to select or behave inconsistent: at one time prefer i to j and at the other time preferring j to i. That is, people are often not sure which alternative they should select, nor do they always make the same choice under seemingly identical conditions. In order to account for the observed inconsistency and the reported uncertainty, choice behavior has been viewed as a probabilistic process. The idea is that there is a pattern to these inconsistencies, and that although the choosing subject is not absolutely consistent, she is still probabilistically consistent.

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

[5] An order $>$ is negatively transitive iff $\forall x, y, z : (x \not> y \wedge y \not> z) \rightarrow x \not> z$.

The general idea is this.[6] Suppose we force subjects to repeatedly make *binary choices* between options i and j under otherwise identical conditions. This could be either a choice what to buy for dinner or, more plausibly, a *perceptual choice*: which of i or j is louder, heavier, more greenish ...? The question then is what should we, as outside observers who know the objective physical properties of i and j, make of subjects' inconsistent choices? In very rough terms, the idea is that we find the consistency somewhere else: we assume that each time the subject makes a choice between i and j the system is shocked systematically, i.e., we assume that our subjects do never actually observe i but rather $i + \epsilon$ where ϵ is a systematic "tremble" drawn from a particular probability distribution. We can then explain subjects' choice behavior as rationally consistent with what they observe, if we factor in that each of their choices is subject to such systematic noise. Depending on the probability distribution of the trembles, we will find that the "mistakes" that subjects make in choosing consistently between i and j depend on the actual values of i and j: for example, if i and j are nearly identical fragrances, the probability of mix-up is higher than when one is Channel No. 5 and the other is the smell of a freshly cooked steak.

Quantal Response Equilibrium. Such stochastic choice models are not confined to perceptual decision making. From work in behavioral game theory it is known that real people are not perfectly rational utility maximizers. The decisions of actual people, when faced with a choice, are similarly noisy, in a way that is nevertheless related to the utility of the options in question. If we assume that the "trembles" with which agents can perceive the quality of their choices are drawn from an extreme-value distribution (roughly: small trembles very frequent, large trembles highly unlikely), then their choice behavior can be modeled by a so-called *logit probabilistic choice rule* which states that the probability $P(i)$ of selecting a given decision i is an exponential function of i's utility u_i (see [20], [21] and [7] for details):

$$P(i) = \frac{\exp(\lambda u_i)}{\sum_j \exp(\lambda u_j)} . \tag{2}$$

Here, λ is a non-negative parameter that measures the degree of rationality of the decision maker. $\lambda = 0$ 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.

The connection to perceptual classification is obvious: when stimuli (such as the agent's own information state, her action choices etc.) are clearly discernible, i.e., when λ is big, the agent will make decisions that are on average more in conformity with hard-edged, classical rationality. But if an agent's perception is error-prone, i.e., when λ is small, her decisions will on average diverge more severely from standard rationality.

This much concerns a single agent. But if the source of imperfection in decision making is systematic, then it may also systematically alter the structure of equilibria

[6] 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 [28] or [19].

that ensue in strategic situations where all players choose with a globally fixed λ.[7] 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 [20,21,8].[8]

Example. Consider 2-state, 2-message signaling game for type-matching with, for simplicity, a uniform prior. 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. This simple example already illustrates the crucial features needed for an account of vagueness. Both the descriptive meanings and the imperative meanings of signals show a number of welcome properties: they place positive probability on all possibilities, and thus do not define a sharply delimited set of types/actions as their meaning. Of course, the simple 2-state case is rather trivial in this respect. But it turns

[7] See [25] for a model that dispenses with the homogeneity of λ among players.

[8] As λ goes to infinity, QREs converge to some NE, the borderline case of perfect rationality.

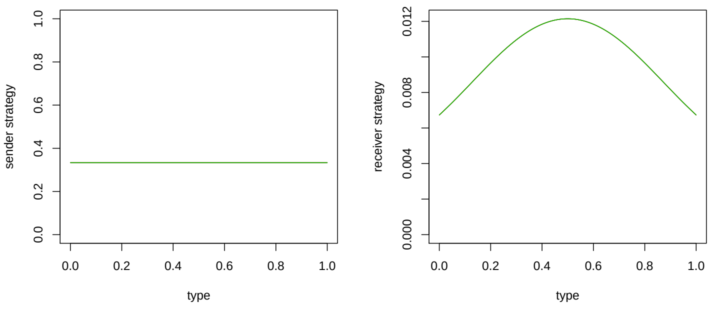

Fig. 2: Babbling equilibrium

out that if we attend to a more interesting case with a topological state space, as described in Section 2, logit equilibria indeed give rise to continuously blended category boundaries of the relevant kind.

This can be shown by simulation. We assumed a game with 100 states that are arranged in the unit interval with equal distances. We considered three signals, and chose the value $\sigma = .2$. For this, simulations show behavior that is similar to the example discussed above. If λ is small, there is only a babbling equilibrium. It is depicted in Figure 2. The left hand side shows the sender strategies. For each type, all three signals are equally likely. The right hand side shows the receiver strategies. Each signal is interpreted as the same probability distribution over types. This distribution is bell shaped and centered at 0.5. 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 neighoring 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. This is as we would like it to be for an account of vagueness, and, as far as we can tell, especially this latter aspect has received little attention so far.

Conclusion. We can thus conclude that 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. The more general reason why this model gives rise to vague meanings is also natural: the sender may only imperfectly observe the state that

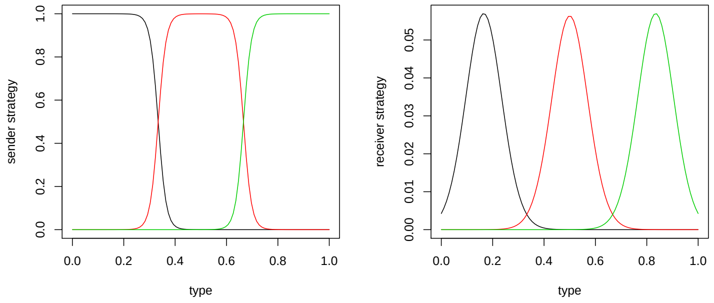

Fig. 3: Separating equilibrium

she wants to communicate, or she may not know whether too much precision is actually needed; similarly for the receiver.

Of course, the QRE raises a number of fair questions too. Even if we accept that all natural language expression 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 with more nuanced topological properties of the space given by T and the utility function U. E.g., what happens if some elements of T are clearly distinguishable from all others, while some others are not? Further issues for future research are to extend the two-agent models to more realistic multi-agent models and to take the step from simulation to more analytic results where possible.

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. van Deemter, K.: Utility and language generation: The case of vagueness. Journal of Philosophical Logic 38(6), 607–632 (2009)
4. Fine, K.: Vagueness, truth and logic. Synthese 30(3–4), 265–300 (1975)
5. Fudenberg, D., Levine, D.K.: The Theory of Learning in Games. MIT Press (1998)
6. Gärdenfors, P.: Conceptual Spaces: The Geometry of Thought. MIT Press (2000)
7. 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)
8. 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)
9. Hendon, E., Jacobsen, Sloth, B.: Fictitious play in extensive form games. Games and Economic Behavior 15(2), 177–202 (1996)

10. Hobbs, J.: Granularity. In: Proceedings of the International Joint Conference on Artificial Intelligence (1985)
11. de Jaegher, K.: A game-theoretic rationale for vagueness. Linguistics and Philosophy 26(5), 637–659 (2003)
12. Jäger, G.: The evolution of convex categories. Linguistics and Philosophy 30(5), 551–564 (2007)
13. Jäger, G., Koch-Metzger, L., Riedel, F.: Voronoi languages (2009), manuscript, University of Bielefeld/University of Tübingen
14. Jäger, G., van Rooij, R.: Language stucture: Psychological and social constraints. Synthese 159(1), 99–130 (2007)
15. 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)
16. Lewis, D.: Convention. A Philosophical Study. Harvard University Press (1969)
17. Lipman, B.L.: Why is language vague? (2009), manuscript, Boston University
18. Luce, D.R.: Semiorders and a theory of utility discrimination. Econometrica 24, 178–191 (1956)
19. Luce, D.R.: Individual Choice Behavior: A Theoretical Analysis. Wiley, New York (1959)
20. McKelvey, R.D., Palfrey, T.R.: Quantal response equilibria for normal form games. Games and Economic Behavior 10(1), 6–38 (1995)
21. McKelvey, R.D., Palfrey, T.R.: Quantal response equilibrium for extensive form games. Experimental Economics 1, 9–41 (1998)
22. Myerson, R.B.: Game Theory: Analysis of Conflict. Harvard University Press (1991)
23. Nosofsky, R.M.: Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General 115(1), 39–57 (1986)
24. Pinkal, M.: Logic and the Lexicon. Kluwer (1995)
25. Rogers, B.W., Palfrey, T.R., Camerer, C.: Heterogeneous quantal response equilibrium and cognitive hierarchies. Journal of Economic Theory 144(4), 1440–1467 (2009)
26. van Rooij, R.: Vagueness and linguistics. In: Ronzitti, G. (ed.) Vagueness: A Guide. Springer (2010)
27. Sainsbury, M.: Is there higher-order vagueness? The Philosophical Quarterly 41(163), 167–182 (1991)
28. Thurstone, L.L.: Psychophysical analysis. American Journal of Psychology 38, 368–389 (1927)
29. Williamson, T.: Vagueness. Routledge (1994)