Available online at www.sciencedirect.com
ScienceDirect
Journal of Mathematical Psychology 51 (2007) 359–382
Journal of Mathematical Psychology
www.elsevier.com/locate/jmp

Evolutionary models of color categorization based on discrimination

Natalia L. Komarova[a,*], Kimberly A. Jameson[b,*], Louis Narens[c]

[a]Department of Mathematics, University of California, Irvine, USA
[b]Institute for Mathematical Behavioral Sciences, University of California, Irvine, USA
[c]Department of Cognitive Sciences, University of California, Irvine, USA

Received 30 June 2006; received in revised form 22 May 2007
Available online 6 August 2007

Abstract

Specifying the factors that contribute to the universality of color categorization across individuals and cultures is a longstanding and still controversial issue in psychology, linguistics, and anthropology. This article approaches this issue through the simulated evolution of color lexicons. It is shown that the combination of a minimal perceptual psychology of discrimination, simple pragmatic constraints involving communication, and simple learning rules is enough to evolve color-naming systems. Implications of this result for psychological theories of color categorization and the evolution of color-naming systems in human societies are discussed.

1. Introduction

1.1. Psychological models of color categorization

An ongoing debate in the psychological study of human color categorization and naming is whether universal tendencies exist in the ways different linguistic societies categorize and name perceptual color experiences. The most popular view in the empirical literature on color categorization and naming is that the commonalities of color categorization across individuals and cultures are largely explained by two factors: (i) physiological features of human perceptual color processing, and (ii) universal features of individual psychological processing believed to underlie color experience. The established position in the area is a strong form of this *universalist* view that asserts that the pan-human uniformity in human visual processing gives rise to a regular, if not uniform, pan-human phenomenological color experience, and that this regularity is the basis for the empirically observed regularity in color categorization across cultures (see Kay, 2005; Kay & Regier, 2003; Regier, Kay, & Cook, 2005; and the references therein).

A very different alternative view is a *relativist* one asserting that very little in the way of "universal tendencies" exist, and that most of the "universalist" findings in the literature are more attributable to constraints imposed by the empirical assessment of the phenomena than they are to actual features of color categorization phenomena (Saunders & van Brakel, 1997).

In addition to these "established" and "alternative" views, a range of perspectives exist that blend the *universalist* and *relativist* approaches to varying degrees, with the aim of providing a comprehensive understanding to the ways different linguistic groups categorize and name their color experience. (For a representative survey of the range of existing perspectives see the edited volumes of *Cross-Cultural Research*, 2005a, 2005b; Hardin, 1988; Hardin & Maffi, 1997; *Journal of Cognition & Culture*, 2005.)

Although a considerable amount of detailed empirical and theoretical research has examined a range of factors influencing the phenomena of color naming, formal models have not emphasized the pragmatic and communication conditions that may be needed for the development and maintenance of a color categorization system shared among humans. In this article, we consider both *intra-individual discrimination* and *inter-individual communication* to be essential for establishing, sharing, and maintaining of a color communication code. We show through mathematical

*Corresponding authors.
E-mail addresses: komarova@math.uci.edu (N.L. Komarova), kjameson@uci.edu (K.A. Jameson).

0022-2496/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2007.06.001

analyses and simulations that the simplest forms of discrimination and communications are sufficient for the evolution of color-naming systems using simple learning algorithms. Although here we focus on the learning and evolution of color categories among artificial agents—and do not investigate human categorization phenomena—the implications of our results on human cross-cultural color-naming research are discussed.

1.2. The signaling environment

Color naming is an example of a signaling system, with a color name signaling a color appearance. The assignment of meaning to signals in simple signaling situations has been studied using evolutionary game theory and evolutionary algorithms. It is an important issue in biology (e.g., Fitch, 2000; Hauser, 1996; Smith & Harper, 2003), artificial intelligence (e.g., Niyogi & Berwick, 1996; Steels & Vogt, 1997), and social evolutionary theory (e.g., Skyrms, 1996). With a few exceptions, the evolutionary algorithms have been applied to situations in which a similarity structure on the meanings of the signals was neither needed nor used. However, for some signaling systems to be effective, a similarity structure on the things to be named is required. Color naming requires such a similarity structure.

Perceptually, colors vary continuously, and the perceptual space of colors is representable as a topology. Human color naming reflects the perceptual topology in various ways. For example, in human color naming, each name describes a portion of color space where each color in the portion can be connected to each other color in the portion via a continuous curve that is completely contained in the portion. As a consequence, in almost all empirical cases, the meaning of a color name will never be a set of colors which can be partitioned into disconnected parts. Colors that are perceptually very similar—that is, colors that are very "close" to one another in the topology—will almost always be described by the same name, while different names will generally denote dissimilar colors. Also, the cognitive organization of the color names will inherit a similar topological organization from the perceptual topology of colors; that is, a cognitive organization of names will emerge that will have a global structure similar to the global structure of the color topology. One goal of this article is to understand from an evolutionary perspective the emergence and stability of such characteristics of color-naming systems. This is done here by investigating two of the simplest yet interesting topologies of colors from the mathematical and naming points of views: colors organized on a line and colors organized on a circle.

From the point of view of human and primate color theory, colors organized on a line and those organized on a circle are natural subspaces to consider for several reasons: First, as a rule, the standard scientific paradigm is to initially model the simplest case from the domain investigated. The color circle continuum together with the continuous line segments of lightness (or brightness) and saturation represent three dimensions widely considered essential for the understanding of perceptual color experience. One widely used representation of human color experience is the Munsell color order system (e.g., Newhall, Nickerson, & Judd, 1943). Fig. 1 depicts the Munsell color solid arrangement from this system. The system aims to model perceptually uniform color differences along the three dimensions illustrated.

In this article, we examine the evolution of color naming for a circular dimension and separately for a linear dimension. The circular dimension is like the hue dimension shown in Fig. 1, and the linear dimension is like the chroma and value dimensions shown in that figure. We base our algorithms on just-noticeable-differences (or *jnds* for short) in color. We identify perceptual colors with physical stimuli, which we call "chips," and we consider them to produce perceptual properties that are structurally similar to those produced by the chips making up the Munsell solid. We consider the separate modeling of these dimensions to be a natural, first step towards modeling the full color solid. As discussed later, extensions of the present modeling methods to the full solid appear to be straightforward.

By basing our investigations on discrimination along linear and circular arrangements of chips, we allow for comparisons with many empirically based human and non-human primate color categorization articles which often use stimuli selected from a portion of the Munsell color solid (e.g., the World Color Survey database; see Matsuno, Kawai, & Matsuzawa, 2006; Matsuzawa, 1985; Regier et al., 2005), and with the closed hue circle continuum

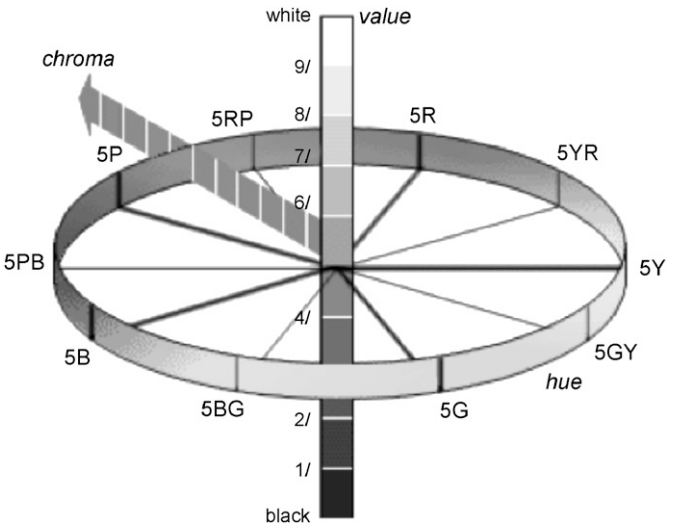

Fig. 1. Munsell book of color system. Dimensions of the Munsell three-dimensional solid consist of color samples (or *chips*) arranged by approximately uniform color differences along each of the three dimensions, (i) a circular dimension of *hue* for representing non-achromatic colors, (ii) a linear dimension of *value* (lightness, or brightness), and (iii) a linear dimension of *chroma* (saturation). Image credit Bruce MacEvoy © 2006. Retrieved 03/15/06 from www.handprint.com. Reproduced with permission.

found in the Farnsworth-Munsell 100-Hue test, which is derived from Munsell book of color stimuli.

Modeling jnds for both circular and linear arrangements of color chips taps into the continuous nature of the perceptual constructs of hue, saturation, and brightness. Jnd variation across different observers naturally exists. For example, "normal" hue discrimination permits perceptual differentiations around the entire hue circle continuum, whereas the hue discrimination of observers with extreme color deficiencies reflects constrained jnd differences on the hue circle in which some discriminations collapse across the hue circle. The latter renders the circle into an elliptical contour that is more closely approximated by a line than a circle, and in the present simulations this line model is used to initially approximate color deficient observers. Modeling both circular and linear continua thus permits the additional opportunity to investigate the consequences of interactions among agents from a heterogeneous population composed of "normal" and deficient observers. As suggested by Jameson (2005a, 2005b), the possibility exists that in heterogeneous populations, category distinctions may be influenced by a need to disambiguate the communication of categories among varying observer types.

Our general approach to learning in the evolutionary modeling of color naming is to start with very simple evolutionary algorithms that are incapable of achieving good categorization, and gently increase the complexity of the algorithm until categorization is achieved. Although many alternatives exist for evolving color lexicons, the simple features and the forms of evolutionary algorithms used here are presumably much less complicated than those found in the evolution of human color categories.

In the following sections, systems of color categorization are evolved in color signaling games. Two classes of learning algorithms are employed: *individual* and *population*. Individual learning is evaluated and updated by comparing the individual's categorization of currently presented stimuli with his previous categorization. In population learning, his categorization of current stimuli is also compared with that of another member of the society as part of the updating. Thus population learning may be looked at as an extension of individual learning that includes features of the categorization behavior of other individuals of the population. In some cases, population learning uses an index (called "fitness") which relates how good at categorizing agents are based on their performance in previous rounds of the game.

Realistically, when a child learns a society's naming system for colors, she is learning an already established system. Here agents evolve a naming system rather than learn an existing system. When a society evolves a naming system, there is not an established system to begin with, but rather a series of different naming systems that reaches an endstate (i.e., an *equilibrium*), where there is no or only very minor deviations over time. Our article studies the possible endstates of such changing or evolving systems. Because both "learning" and "evolving" have been used in the literature to describe the time course of strategies used by interacting individuals in achieving their goals, a clear distinction between the two terms cannot be made while accounting for the literature.[1]

Our algorithms provide criteria to evaluate whether some color signaling behaviors might be more successful as categorization behaviors than others. This provides for a selection pressure which in turn influences evolving signaling strategies, and an agent produces a sequence of signaling systems involving color categories and names. One way such sequences are produced is through forms of reinforcement learning. These forms of reinforcement have the following property:

> Each element of the sequence, except for the first, tends to be more successful in naming than it predecessor, except after some point in the sequence, where it tends to be just as successful in naming as its predecessor.

Simulations provided in Sections 3 and 4 show that for reinforcement learning, the resulting sequence of color signaling systems reaches a limit that, for all practical purposes, can be considered a stable categorization system for the naming of colors.

Human languages categorize colors in a variety of ways. This has produced a diversity of explanations from historians, linguists, anthropologists, psychologists, and physiologists regarding the observed regularities found in the naming of colors from ancient languages and different ethnolingustic populations. The dominant view in the literature for explaining these regularities uses the six Hering primaries—white, black, red, yellow, green, and blue—as the foundation for explaining the commonalities found in color naming (see discussion in Jameson, 2005c). Many researchers have gone further and tried to explain the observed commonalities in terms of the physiological opponent processing of color in the peripheral visual system. Such a view suggests that the visual processing system assigns privileged status to the Hering primaries. This physiological explanation has been widely employed to provide a theoretical basis for pan–human naming regularities based on presumed physiological color processing universals. While still prevalent in the literature, this physiological approach to color naming has been abandoned by most of the major researchers in the area. For example, Boynton (1997) wrote,

> … I would argue that all eleven basic colors are perceptual fundamentals, and that the concept of fundamental neural responses, as defined by Kay,

[1]There is also sometimes a difference in the kind of algorithms used in learning and evolutionary studies. But this is not an important factor here, because (i) this article's positive results are based on dynamics produced by reinforcement algorithms, which are used both in learning and evolutionary studies, and (ii) mathematical results about Darwinian-like algorithms and reinforcement learning algorithms show that they are closely connected (e.g., Beggs, 2005; Börgers & Sarin, 1997).

> Berlin, and Merrifield (1991), should be expanded to include all eleven. Their appeal to the early research of DeValois and his colleagues [the suggested hardwired neural basis of 'red,' 'green,' 'yellow,' and 'blue' experiences] is misguided, if only because sensations surely do not arise from the lateral geniculate nucleus, which is the site of their recordings. (p. 148)

Of course, the six Hering primaries may continue to be considered as a universal basis for color naming for reasons other than physiological opponent color processing, as for example in the empirically based approach of Regier et al. (2005). They write,

> It is widely held that named color categories in the world's languages are organized around universal focal colors and that these focal colors tend to be chosen as the best examples of color terms across languages. However, this notion has been supported primarily by data from languages of industrialized societies. In contrast, recent research on a language from a nonindustrialized society has called this idea into question. We examine color-naming data from languages of 110 nonindustrialized societies and show that (i) best-example choices for color terms in these languages cluster near the prototypes for English *white*, *black*, *red*, *green*, *yellow*, and *blue*, and (ii) best-examples choices cluster more tightly across languages than do the centers of category extensions, suggesting that universal best examples (foci) may be the source of universal tendencies in color naming. (p. 8386)

However, more recently Regier, Kay, and Khetarpal (2007) have adopted a different approach:

> The nature of color categories in the world's languages is contested. One major view holds that color categories are organized around universal focal colors, while an opposing view holds instead that categories are defined at their boundaries by linguistic convention. Both of these standardly opposed views are challenged by existing data. Here, we argue for a third view, originally proposed by Jameson and D'Andrade: that color naming across languages reflects optimal or near-optimal divisions of an irregularly shaped perceptual color space. We formalize this proposal, test it against color naming data from a broad range of languages, and show that it accounts for universal tendencies in color naming, while also accommodating some observed cross-language variation. (p. 1436)

Similar to Regier et al. (2007), the present evolutionary approach to color naming is also based on Jameson and D'Andrade (1997) and that view as extended by Jameson (2005d). Jameson and D'Andrade (1997) and Jameson (2005d) suggest that similar cross-cultural color naming arises from widespread human tendencies to evolve color categorization systems which aim for optimal partitions on color differences that are non-uniformly distributed within an irregularly shaped perceptual color space.[2] Their theory also emphasizes that the evolution of human categorization systems is additionally constrained by information processing demands in the course of effectively capturing such perceptual non-uniformities and irregularities. This emphasis differs from the strict perceptual salience emphasis often seen in the literature (e.g., Hardin, 2005; Kay, 2005; Kuehni, 2005). In addition to perceptual considerations, Jameson (2005d) proposes further constraints on human color categorization arising from cognitive emphases such as polar opposition, symmetry preference, and connected-set features in developing category systems (also see Garner, 1974); as well as constraints from socially dependent pragmatic communication features. Jameson and D'Andrade's *interpoint distance model* (IDM) thus proposes that perceptual, cognitive and socio-cultural features all figure prominently in human color category system evolution (Jameson, 2005d).

At this initial juncture, the methods formulated here do not permit examination of many of the human color-naming principles described in Jameson (2005d), because, for example, as a prudent first step it was necessary to constrain our investigations to societies of homogeneous agents and—except for one case—homogeneous features of color space. Nevertheless, the portion of our approach described in Section 5 validates features of the Jameson (2005d) theory which state that in a regularized space of uniform color differences, effective communication of color categories should produce color categories that are connected regions of approximately equal size. Extensions of these results to non-homogeneous color spaces and non-homogeneous populations of agents are discussed in Section 5.

2. Mathematical framework for modeling categorization

2.1. Color stimulus domain and the definition of categorization

We now describe investigations of categorization on two sets of color stimuli, one organized along a line, and the other along the continuum of a circle. These sets consist of discrete arrays of color chips arranged according to perceptually jnds; that is, arranged so that adjacent chips are at thresholds of perceptual discriminability.

Suppose an array contains n color chips, $1,\ldots,i,\ldots,n$, such that i and $i+1$ are adjacent, no other adjacencies occur if the chips are on a line, and n is adjacent to 1 if the chips are on a circle. For such arrays we give the following mathematical definition of color categorization:

As the first step, let us consider any two possible categories, say, "light" and "dark." A *categorization* is a mapping from $\{1,\ldots,n\}$ to the interval $[0,1]$, which can be

[2] An approach to optimal color categorization that uses very different assumptions and algorithms, and reaches different conclusions from this article, is presented by Griffin (2006).

presented as an n-tuple of numbers, $F = (f_1, \ldots, f_n)$ with real numbers $f_i \in [0, 1]$. If a color i is presented to a viewer, the viewer with categorization F will assign the label "light" to color i with probability f_i, and the label "dark" with probability $1 - f_i$. F is said to be *probabilistic* if and only if at least one of the entries f_i is not equal to 0 or 1, and *deterministic* if and only if $f_i \in \{0, 1\}$ for all i, $1 \leqslant i \leqslant n$.

Note, first, we do not specify how "light" and "dark" categories initially arise—they could arise for any number of exogenous or endogenous reasons. Second, any other initial categorical distinction (e.g., yellow and blue, warm and cool) will serve equally well for describing this kind of category learning. Our algorithms will assume for this case that a demand exists at the level of the individual agent to differentiate "light" from "dark," which initiates category learning at the individual level. Such an assumption appears realistic—the need to specify a category system on a domain follows from demands for domain differentiation. And it is realistic in practice: for example, human categorization systems that exhibit only 2 color categories are known to exist. In addition, the existence of a universal human tendency to cognitively organize stimulus domains using polar opposition, or symmetry, could also initiate this form of binary categorization (Garner, 1974; Jameson, 2005d, 2005e; Smith & Sera, 1992). Alternatively, pragmatic needs—such as the need to specify the color of a valued food source as differing from other resources—may also serve as pressure to initiate categorization. At this point we only need to accept that some unspecified pressure to make a categorical distinction presents itself to the individual agent.

Next, we extend this notion to multiple color categories. Suppose we have m categories. We denote the m categories by $1, \ldots, m$. (When it is not clear by context when a number i denotes a color chip or a category, it will be made explicit by saying, for example, "the color i," or "the category i.") A *categorization* is a vector function from $\{1, \ldots, n\}$ to $[0, 1]^m$. We can represent it as an n-tuple of vectors, $\mathbf{f}_1, \ldots, \mathbf{f}_n$, where $\mathbf{f}_i$ are m-dimensional vectors whose non-negative entries sum up to 1 (i.e., they belong to an m-simplex). The notation $[\mathbf{f}_i]_k$ is to be interpreted in the following manner:

$$[\mathbf{f}_i]_k = Prob(\text{the agent assigns the category } k \text{ to the color chip } i). \quad (1)$$

A *deterministic categorization* is an n-tuple of integers, $(F_1, \ldots, F_n)$, where F_i is in $\{1, \ldots, m\}$; that is, a deterministic categorization assigns to each color chip exactly one of the categories, $1, \ldots, m$.

2.2. The range of similarity

Our categorizing algorithms are based on the following idea: Colors that are highly similar perceptually to one another are highly likely to belong to the same category. More specifically it is based on the following three principles: (i) categorization is important; (ii) to be useful, categorization should attempt to minimize ambiguity, and (iii) when color is a salient or meaningful cue for categorization, two randomly chosen objects that have similar color appearances are more likely to be categorized together than are two objects that have dissimilar colors. These three principles are summarized in the concept of a *similarity range*, denoted k_{sim}.

By definition, k_{sim} is the minimum difference between the color chips for which it becomes important to treat them for *pragmatic purposes* (and not for *perceptual purposes*) as different color categories. The parameter k_{sim} is defined to be a fixed integer, $0 \leqslant k_{\text{sim}} \leqslant n - 1$, where n is the number of color chips. Pragmatically speaking (see principle (i) above), it pays off to assign colors outside the k_{sim}-range to different color categories (principle (ii)), and colors within the k_{sim}-range to the same color category (principle (iii)).

k_{sim} is interpreted as being related to the utility of color categorization and is defined by the environment and the lifestyle of the individual agents. It is used to reflect the notion of the pragmatic color similarity of the chips. For instance, suppose one individual shows another a fruit and asks him/her to bring another fruit "of the same color." It is a nearly impossible task to bring a fruit of a color perceptually identical to the first, because different lighting, different color background and slight differences in fruit's ripeness contribute to its perceived color. Therefore, to satisfy "of the same color" of a fruit's ripeness in practical terms, the individual must be able to ignore such unimportant perceptual differences and bring a fruit that is "of the same color" practically.[3] It may also be just as important to be able to distinguish ripe, edible, "red" fruit from the unripe, "green" ones. The parameter k_{sim} is intended to set a scale at which color differences become important in the everyday world. It tells us that most of the time, certain objects with colors within the k_{sim} range will have similar pragmatic properties which they will not have with larger color differences.

It is important to emphasize that the range k_{sim} is not another perceptual version of "just noticeable difference." Colors that differ by a perceptual jnd are well within the similarity range, as are adjacent chips in the Munsell color system described earlier. In general, some colors within the k_{sim} range are easily distinguished perceptually by any agents. The notion of k_{sim} intends to capture the *importance* of categorizing two chips as being "different" rather than "the same."

We will shortly present a simple model which expresses the evolutionary importance of color categorization with k_{sim} being a constant number. The next level of complexity considered in this article introduces a non-homogeneous $k_{\text{sim}}^{(i)}$, that is, it includes the ability to make different fine-color distinctions in naming in different portions of the color stimulus domain. Both these are special cases of a

[3] The challenge of modeling novel stimulus classification is discussed again in Section 5.1.2.

general notion of similarity probability, in which there is an expected likelihood, $P_{ij} = P_{ji}$, that objects of colors j and i have similar properties, pragmatically speaking. In other words, instead of a scalar quantity k_{sim}, in a more general case we could talk about a similarity matrix. The special case of the homogeneous k_{sim} then follows if we set $P_{ij} = 1$ for $|i-j| \leqslant k_{\text{sim}}$ and $P_{ij} = 0$ otherwise.

Note that mathematically speaking, it is possible to perform a limiting transition to a continuous description of the stimulus domain. To do this, we need to relax the assumption that the distance between two neighboring chips is a jnd, and instead to tend this distance to zero. As a result, the number of color chips will tend to infinity, $n \to \infty$. The similarity range, k_{sim}, will change proportionally. The number n (the perceptual resolution of the color chips) can be arbitrarily high, and no results presented in this article depend on the actual number n. What is important is the relative (to n) size of the similarity range, which is basically how many intervals of length k_{sim} span the investigated color stimulus domain.

We now consider the evolution of categories for an individual agent based only on his personal color experience.

2.3. Two types of discrimination games and their success rates

The *discrimination game* is defined as follows: Two color chips i and j are chosen from a distribution and presented to a simulated agent, or *viewer*. The viewer classifies them according to F, that is, he assigns labels v_i and v_j, respectively, to i and j, where v_i and v_j are in $\{1, \ldots, m\}$. The discrimination game is said to be *solved successfully* if and only if for $|i-j| > k_{\text{sim}}$, the viewer chooses $v_i \neq v_j$. In other words, if the color chips i and j are sufficiently different (i.e., $|i-j| > k_{\text{sim}}$), then the viewer assigns them to different categories. The discrimination game is said *to fail* if and only if for $|i-j| > k_{\text{sim}}$, the viewer chooses $v_i = v_j$; that is, if the color chips are sufficiently different, then the viewer assigns them to the same category. If the chips i and j are within their similarity range (i.e., $|i-j| \leqslant k_{\text{sim}}$), they are discarded and the game is not played.

The *discrimination–similarity game* includes the data about chips that are within of k_{sim} of each other. We say that two chips i and j are k-similar if the inequality $|i-j| \leqslant k$ holds. The discrimination–similarity game is defined as follows: If the two selected chips are further apart than k_{sim}, then the game is a success if the viewer puts them into different categories. For two chips within distance k_{sim}, the game is a success if the viewer assigns them to the same category.

The *success rate* for games is defined as follows: For a fixed distribution, pairs of color chips are given to a viewer playing either the discrimination game, or the discrimination–similarity game. After M rounds of the game, the fraction of successful games is denoted as S_M. The success rate S is then given by the equation $S = \lim_{M\to\infty} S_M$. This quantity is equivalent to the probability of having a successful round of the game.

Finally, we define an *optimal categorization* for a specific game as a categorization that maximizes the success rate for this game.

2.4. Optimal categorizations for discrimination games

Let us suppose the following:

- $m = 2$ (2 color categories),
- n is even,
- $0 < k_{\text{sim}} < n/2$,
- the color chips are chosen from a uniform distribution, and a number of rounds of discrimination game are played.

Statement 1. Suppose the four assumptions above hold. If the chips have a linear arrangement, then the following two categorizations maximize the success rate:

(1) $F_i = 2$ for $i \leqslant n/2$ and $F_i = 1$ for $i > n/2$, and
(2) $F_i = 1$ for $i \leqslant n/2$ and $F_i = 2$ for $i > n/2$.

If the chips have a circular arrangement, then the following categorization maximizes the success rate:

$$F_i = 2 \quad \text{for } i \leqslant n/2 \quad \text{and} \quad F_i = 1 \quad \text{for } i > n/2.$$

Also, any shift of the categorization pattern along the circle will also maximize the success rate. There are no other categorizations with success rates equal to or larger than those achieved by the above categorizations.

Proof. Let us denote the quantity $\|i-j\| \equiv d_{ij}$, the distance between chips i and j. Note that in the case of a circular arrangement we take the shortest distance along the circle.

In general, the probability of success (the success rate) of a discrimination game is given by the following:

$$S = W^{-1} \sum_{j=1}^{n} \sum_{\|l-j\| > k_{\text{sim}}} v_j v_l (1 - \delta_{F_j, F_l}),$$

where v_j is the probability to draw the color chip j, and the categorization is defined by $F_j \in \{1, 2\}$. We also used the notations:

$$W = \sum_{j=1}^{n} \sum_{\|l-j\| > k_{\text{sim}}} v_j v_l \quad \text{and} \quad \delta_{xy} = \begin{cases} 1, & x = y, \\ 0 & \text{otherwise}, \end{cases}$$

where δ_{xy} is the Kronecker symbol. Setting $v_j = 1/n$, we obtain

$$W = 1 - \frac{2k_{\text{sim}} + 1}{n}$$

in the circular geometry and

$$W = \left(1 - \frac{k_{\text{sim}}}{n}\right)\left(1 - \frac{k_{\text{sim}} + 1}{n}\right)$$

in the linear geometry. In both cases, to maximize the success rate, it is enough to maximize the quantity

$$\bar{S} = \sum_{j=1}^{n} \sum_{\|l-j\|>k_{\text{sim}}} (1 - \delta_{F_j,F_l}).$$

This is reminiscent of the Hamiltonian in a one-dimensional Ising model with non-local interactions.[4]

Let us consider all categorizations where the number of chips with $F_j = 1$ is A and the number of chips with $F_j = 2$ is $n - A$ (here A is some integer, and without loss of generality we can assume that $0 \leqslant A \leqslant n/2$). Now, the quantity $\bar{S}$ can be rewritten as

$$\begin{aligned}\bar{S} &= \sum_{j=1}^{n}\sum_{l=1}^{n}(1 - \delta_{F_j,F_l}) - \sum_{j=1}^{n} \sum_{\|l-j\|\leqslant k_{\text{sim}}} (1 - \delta_{F_j,F_l}) \\ &= 2A(n-A) - \sum_{j=1}^{n} \sum_{\|l-j\|\leqslant k_{\text{sim}}} (1 - \delta_{F_j,F_l}). \end{aligned} \quad (2)$$

The first term in this expression does not depend on the particular configuration but only on the number of chips of each color (that is, on A). The second term has to be minimized over all configurations. This term contains "interactions" of a chip j with all chips l such that $d_{lj} \leqslant k_{\text{sim}}$. The meaning of this term is penalizing color differences: every time the two chips in a k_{sim}-neighborhood are categorized differently, $\bar{S}$ acquires a negative contribution. Minimizing this "penalty" term is equivalent to maximizing

$$U = \sum_{j=1}^{n} \sum_{\|l-j\|\leqslant k_{\text{sim}}} \delta_{F_j,F_l}, \quad (3)$$

that is, maximizing the number of chips of the same color category within neighborhoods of size k_{sim}. The configuration which corresponds to the maximal number of chips of the same category in k_{sim}-neighborhoods is the one with the minimum number of boundaries, or transitions from one category to the other. To see this, let us rewrite expression (3) as a sum of two terms:

$$U = \sum_{F_j=1} \sum_{\|l-j\|\leqslant k_{\text{sim}}} \delta_{F_j,F_l} + \sum_{F_j=2} \sum_{\|l-j\|\leqslant k_{\text{sim}}} \delta_{F_j,F_l}. \quad (4)$$

The first term on the right-hand side corresponds to summing over all chips categorized as color 1, and the second term—to that with all chips categorized as color 2. It is easy to see that the first term is maximized if all the chips with $F_j = 1$ form a patch, such that they are all adjacent to each other (and to one of the boundaries, in the case of a linear geometry). Such a configuration happens to maximize the second term on the right-hand side of Eq. (4). Therefore, the best possible categorization will consist of only two "patches": a patch of color 1 and a patch of color 2.[5]

Next, we need to consider all categorizations with the "2-patch" structure, characterized by the parameter A. We will show that among such categorizations, the one with $A = n/2$ corresponds to the maximum success rate. Let us evaluate the quantity $\bar{S}$. We will use a circular geometry. This reasoning can also be extended to a linear geometry. It is possible to show that for $0 < k \leqslant A$ we have

$$\sum_{j=1}^{n} \sum_{\|l-j\|=k} \delta_{F_j,F_l} = 2n - 4k,$$

for $n/2 > k > A$ we have

$$\sum_{j=1}^{n} \sum_{\|l-j\|=k} \delta_{F_j,F_l} = 2(n - 2A),$$

and for $k = 0$ we have

$$\sum_{j=1}^{n} \sum_{\|l-j\|=0} \delta_{F_j,F_l} = n.$$

Therefore, for $A \geqslant k_{\text{sim}}$ we have

$$\begin{aligned}\bar{S} &= C + 2A(n-A) + n + \sum_{k=1}^{k_{\text{sim}}}(2n - 4k) \\ &= C + 2A(n-A) + D,\end{aligned}$$

where $C = -\sum_{j=1}^{n} \sum_{\|l-j\|\leqslant k_{\text{sim}}} 1$ and $D = n(1 + 2k_{\text{sim}}) - 2k_{\text{sim}}(k_{\text{sim}} + 1)$ do not depend on A. In this case the function $\bar{S}_A$ is maximized by the value $A = n/2$.

For $A < k_{\text{sim}}$ we obtain

$$\begin{aligned}\bar{S}_A &= C + 2A(n-A) + \sum_{k=1}^{A}(2n - 4k) + \sum_{k=A+1}^{k_{\text{sim}}} 2(n - 2A) \\ &= C + 2k_{\text{sim}}n + n + 2A(n - 2k_{\text{sim}} - 1). \end{aligned} \quad (5)$$

Now, we have $d\bar{S}_A/dA = 2(n - 2k_{\text{sim}} - 1) > 0$ (because $k_{\text{sim}} \leqslant n/2 - 1$). Therefore, the optimal A is given by its maximal admissible value, $k_{\text{sim}} - 1$. A direct comparison of the two cases shows that for all $k_{\text{sim}} \leqslant n/2 - 1$, the success rate is maximized by the case $A > k_{\text{sim}}$, and thus the optimal categorization structure is given by $A = n/2$.

This means that in order to achieve an optimal categorization, the number of chips of colors 1 and 2 must be equal. This completes the proof of Statement 1. □

Let us calculate the success rate of the optimal categorization for the uniform sampling of color chips. Assume that chips are arranged in a circle. If the probability to draw any given chip is $1/n$, then the probability to draw a pair of chips with distance d between

[4] Indeed, if the spin values $\sigma_j \in \{-1, +1\}$, then we have

$$\bar{S} = \sum_{j=1}^{n} \sum_{\|l-j\|>k_{\text{sim}}} \left(\frac{1 - \sigma_j\sigma_l}{2}\right) = const - 1/2 \sum_{j=1}^{n} \sum_{\|l-j\|>k_{\text{sim}}} \sigma_j\sigma_l.$$

[5] To be precise, if $A < k_{\text{sim}}$, then there is a whole class of optimal categorizations which includes all configuration where the largest distance between any two chips with $F_j = 1$ is no more than k_{sim}; the two-patch categorization described here belongs to this class.

them, v_d, is given by

$$v_d = \frac{1}{n/2 - k_{\rm sim} - 1/2} \equiv \tilde{v}, \quad k_{\rm sim} < d < n/2, \quad v_{n/2} = \tilde{v}/2.$$

For all pairs of distance d, the probability to belong to two different categories is $2d/n$. The success rate is given by

$$2 \sum_{d=k_{\rm sim}+1}^{n/2} v_d \frac{d}{n}.$$

Therefore, the optimal success rate for the discrimination game is

$$S = \frac{(n/2)^2 - k_{\rm sim}(k_{\rm sim}+1)}{n(n/2 - k_{\rm sim} - 1/2)}.$$

In particular, if $k_{\rm sim} = 0$, then we have $S = \frac{n}{2(n-1)}$: the value of S increases monotonically with $k_{\rm sim}$, reaching $S = 1$ for $k_{\rm sim} = n/2 - 1$.

Next, let us extend our consideration to the case of several color categories. We suppose that

- $m \geqslant 2$;
- n is divisible by m;
- $0 < k_{\rm sim} < n/m$,
- the color chips are chosen from a uniform distribution, and rounds of discrimination game are played.

Statement 1′**.** Suppose the four assumptions above hold. If the chips are points on an interval, then the following categorization maximizes the success rate: chips

$$\frac{(i-1)n}{m} + 1, \frac{(i-1)n}{m} + 2, \ldots, \frac{in}{m}$$

belong to color category v_i, where $1 \leqslant i \leqslant m$ and for all $i \neq j$, $v_i \neq v_j$, see Fig. 2(a). If the chips are arranged on a circle, then the following categorization maximizes the success rate: the circle is divided into regions

$$\left[1, \frac{n}{m}\right], \left[\frac{n}{m} + 1, \frac{2n}{m}\right], \ldots, \left[\frac{(m-1)n}{m} + 1, n\right].$$

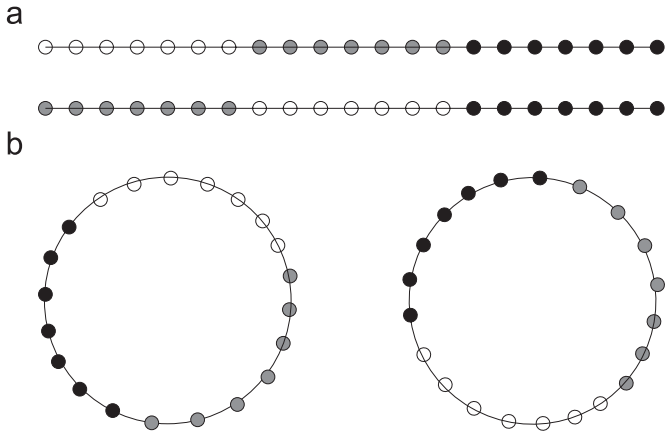

Fig. 2. Some optimal categorizations for (a) an interval and (b) a circle. White, gray and black dots represent the $m = 3$ color categories, and $n = 21$.

All the vectors $\mathbf{f}_i$ with i inside the same region are identical to each other and have only one non-zero component. Any vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ with i and j belonging to different regions are different from each other. Also, any shift of this pattern along the circle maximizes the success rate, see Fig. 2(b). There is no other categorization that has the success rate equal or bigger than those achieved by the above categorizations.

In other words, the most successful categorization deterministically partitions the colors into categories of equal size such that each category is a connected set (as suggested for perceptually uniform spaces by Jameson, 2005d, 2005e). The proof of Statement 1′ is omitted here.

Statements 1 and 1′ above hold if the probability distribution to draw a chip is uniform. This is an important assumption. In the example of Fig. 3, we can see that a non-uniform sampling distribution can break the symmetry (the translational invariance) of the optimal categorizations in a circular geometry. Indeed, let us suppose that each of the chips marked by "X" in the figure is sampled with probability p_1, whereas the rest of the chips are all sampled with probability $p_2 > p_1$ (we have $6p_2 = 1 - 8p_1$). Then, in the limit where $p_1 \to 0$, categorization Fig. 3(a) with $k_{\rm sim} = 2$ gives the success rate of $S_{\rm a} = 1$. The categorization of Fig. 3(b) has a smaller success rate of $S_{\rm b} = 5/8$. For finite values of $p_1 < p_2$, we will still have $S_{\rm a} > S_{\rm b}$. Thus, non-uniformities in the sampling can lock the position of the boundaries of color categories.

In the absence of the uniform sampling distribution, one can also come up with examples of optimal categorizations which do not have the structure described in Statement 1′. Indeed, let us suppose that $n = 12$, $k_{\rm sim} = 3$, $m = 3$, $v_5 = 1$ and $v_d = 0$ for $d \neq 5$. Then the following categorization assures the success rate $S = 1$: chips $1, \ldots, 12$ are assigned color categories

$$1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3. \tag{6}$$

Finally, we note that Statements 1 and 1′ assume that the total number of chips is divisible by m. If this is not the case, then there still exists a family of categorizations that

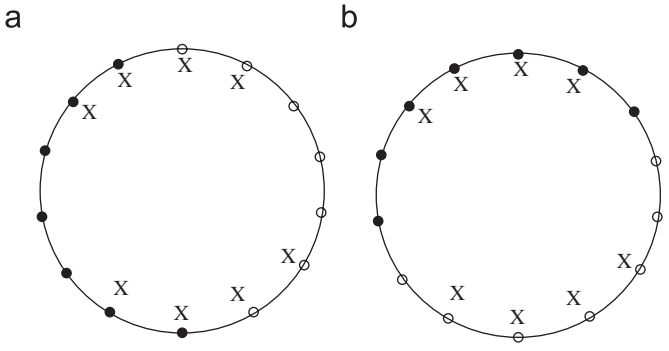

Fig. 3. Symmetry breaking caused by inhomogeneous sampling. Here $n = 14$ and $k_{\rm sim} = 2$. Chips marked by "X" are chosen much less often than the rest. Then (in the limit of zero sampling frequency of the "X"-chips) categorization (a) yields a success rate of 1, whereas categorization (b) yields a success rate of 5/8.

are optimal. They contain all (connected) partitions of the interval (circle) into m regions of maximally equal length.

2.5. Optimal categorizations for discrimination-similarity games

The main difference between the discrimination game (described above) and the discrimination–similarity game is the following: for varying m, the optimal solution of the discrimination game is $m = n$, where each chip has its own category. This leads to the 100% success rate of the discrimination game. Now, if we keep this categorization and play the discrimination–similarity game with $k_{\text{sim}} > 1$, the success rate of this game will be low, because two neighboring chips will always be assigned to different color categories. Therefore, there must be some restriction on the number of categories allowed as an optimal categorization. In fact, the optimal categorization for $k_{\text{sim}} > 0$ in the discrimination–similarity game will include only $m < n$ categories.

In other words, the number of categories m that are learned by a discrimination–similarity learner is smaller than n, the number of chips in the stimulus domain. It is defined by the similarity range, k_{sim}, in such a way that each category must be wider than k_{sim}.

Again, these statements about the optimal categorization imply the assumption that the chips are drawn according to the uniform distribution. For different distributions, the results may change. For instance, if we have $n = 12$, $k_{\text{sim}} = 3$, $m = 3$, $v_3 = 1/2$ and $v_5 = 1/2$, then the categorization system shown in (6) above yields a success rate of 1.

Let us calculate the number of categories in an optimal categorization, under the assumption of the uniform sampling of color chips. Consider the case of a circle of length n, with $k_{\text{sim}} < n/2$. Assume the categorization splits it into m equal regions of length l (such that $n = ml$). Pick two chips at random and play a discrimination–similarity game. What is the chance that this game is successful?

The quantity $d \equiv \|i - j\|$, the shortest distance between two chips along the circle, varies in the interval $1 \leqslant d \leqslant n/2$. We consider two cases: (i) $k_{\text{sim}} < l < n/2$ and (ii) $1 \leqslant l \leqslant k_{\text{sim}}$.

(i) If n is even, we first need to consider the special case of $d = n/2$. There are $n/2$ pairs of this kind, each of them yielding a successful game. For odd n, this case does not enter the argument. Next, consider the case where $1 \leqslant d < n/2$ (for any value of n, even or odd). Let us calculate how many pairs, out of the n possible pairs of distance d, yield successful games. If $l \leqslant d \leqslant n/2$, then all such pairs belong to different categories, which means success, yielding n successful games. If $k_{\text{sim}} < d < l$, then only $n - m(l - d)$ games are successful. Indeed, for each color, we have $l - d$ pairs belonging to it (each of which lead to a failure). Finally, for pairs of size d such that $1 \leqslant d \leqslant k_{\text{sim}}$, the number of successful games is $m(l - d)$, which is equal to the number of pairs of the same color. Let us denote the frequency with which we draw a pair of size d by v_d. Then we have the total success rate,

$$S = \frac{1}{n}\left\{\sum_{d=1}^{k_{\text{sim}}} v_d m(l-d) + \sum_{d=k_{\text{sim}}+1}^{l-1} v_d(n - m(l-d)) + \sum_{d=l}^{n/2-1} v_d n\right\} + v_{n/2}. \tag{7}$$

For the uniform distribution, the probability to draw any pair with distance $d < n/2$ is given by

$$v_d = \frac{1 - v_{n/2}}{n/2 - 1} \equiv \bar{v},$$

where $v_{n/2} = 0$ if n is odd and $v_{n/2} = \bar{v}/2$ if n is even. We have for the uniform sampling,

$$S = \frac{\bar{v}}{n}\left\{\sum_{d=1}^{k_{\text{sim}}} m(l-d) + \sum_{d=k_{\text{sim}}+1}^{l-1} (n - m(l-d)) + \sum_{d=l}^{n/2-1} n\right\} + v_{n/2}.$$

Remembering that $m = n/l$, we obtain

$$S = \frac{\bar{v}}{2l}(2k_{\text{sim}}(l-1) + l(n-l-1) - 2k_{\text{sim}}^2) + v_{n/2}.$$

Differentiating this with respect to l, we find that the extremum is at the point

$$l_c = \sqrt{2k_{\text{sim}}(k_{\text{sim}}+1)}.$$

This is a maximum since

$$d^2S/dl^2 = -\frac{2\bar{v}k_{\text{sim}}(k_{\text{sim}}+1)}{l^3} < 0.$$

(ii) In this case, a similar argument shows that

$$S = \frac{1}{n}\left\{\sum_{d=1}^{l-1} v_d m(l-d) + \sum_{d=k_{\text{sim}}+1}^{n/2-1} v_d n\right\} + v_{n/2}.$$

For the uniform distribution, we get

$$S = \frac{\bar{v}}{2}(n + l - 3 - 2k_{\text{sim}}) + v_{n/2},$$

which is obviously maximized by $l = k_{\text{sim}}$.

Let us now compare the success rates in case (i) with $l = l_c$ and in case (ii) with $l = k_{\text{sim}}$. We have

$$S(l = l_c) - S(l = k_{\text{sim}}) = \bar{v}\left(1 + \frac{3k_{\text{sim}}}{2} - \sqrt{2k_{\text{sim}}(k_{\text{sim}}+1)}\right) > 0.$$

Therefore, the optimal value of l is l_c, and the optimum number of color categories is given by

$$m_c = n[2k_{\text{sim}}(k_{\text{sim}}+1)]^{-1/2}. \tag{8}$$

Note here that computer simulations show that if the value m_c is not an integer, an optimal categorization will divide the n chips in almost equal groups of the length maximally close to m_c.

To summarize, we have defined the number m_c (Eq. (8)), which can be used as an estimate for the actual number of categories in an optimal categorization, given k_{sim} and n. Note that in the case of non-uniform sampling, we can still calculate the optimal number of color categories (assuming that they are of equal size, that is, have the structure described in Statement 1′). For that we need to use the actual distribution values of the pair sizes in formula (7).

The next necessary step is the dynamics of the acquisition of color categorization systems. This is done in the next section by investigating several different learning algorithms. We will present numerical tests which are consistent with the analytical result above. We will show that the optimal number of color categories can be reached as a result of learning dynamics. Namely, if we start with a total number of categories bigger than m_c, and apply a learning algorithm, some of the categories will be weeded out (or reduced to low levels) to match the number of active categories with m_c. Conversely, if we start from a number of active categories smaller than m_c and introduce some additional categories at low levels, the additional categories will eventually be adopted, to arrive at an optimal categorization with m_c color categories.

3. Individual learning

The present research investigated three kinds of individual learning strategies. They were selected because they produce color categorizations that vary in terms of stability and other dynamical features for the games considered. The variability is used to gauge what learning properties and games are related to color categorization systems that structurally resemble those found in cross-cultural studies of human color naming.

Individual learning is concerned with one learner playing a number of games. The learner starts with some initial categorization function F. He is presented with a pair of color chips, (i,j), and assigns categories $v_{i,j}$, respectively, to the color chips i and j. It is assumed that after each game, the individual receives a feedback on the result of the game (i.e., "success" or "failure"). The categorization is updated based on this information. The learner's task is to maximize his success rate.

The update rule is given by an algorithm. Several such algorithms are considered in the next subsection.

3.1. Learning algorithms

In addition to specifying categorization on a stimulus continuum, similarity ranges, and optimal categorizations for two types of discrimination games, we also define three different ways agents learn to categorize color.

3.1.1. A memoryless learner

This learner only performs deterministic categorization; that is, a categorization where the entries of the vectors $\mathbf{f}_i$ are zeros and ones. If the game is a success, the learner stays with the current categorization. If the game fails, the learner chooses a different vector $\mathbf{f}_i$ for one or both chips involved in the game. The choice of the non-zero entry of the new vector(s) is random.

In other words, if the game is a success, then there is no change, and if the game fails, the category of the chip(s) involved is switched at random. In particular, a memoryless learner can go back to a categorization that has already been tried and rejected (thus the name of this strategy). We examined this learning strategy because it has been extensively used for modeling language learning in artificial intelligence type problems, as well as in modeling the evolution of language (Niyogi, 1998; Nowak, Komarova, & Niyogi, 2002).

3.1.2. A smoothing learner

Again, this learner only performs deterministic categorization. In cases where the game is a success, the learner does not change the categorization. If the game fails, she updates the vector $\mathbf{f}_i$ for one or both chips involved in the game by assigning $\mathbf{f}_i = \mathbf{f}_{i+1}$. The maximum number of categories is $m = n$; that is, every chip belongs to a different category.

In other words, if the game fails, the category of the chip(s) involved is made equal to that of its (their) neighbor on the right. We can also use neighbors on the left or a random neighbor.

3.1.3. Reinforcement learner I

These learners allow for stochastic categorizations. Each color chip i is associated with a vector $\mathbf{X}_i$ whose m components are integer numbers that add up to some constant, L: $\sum_{j=1}^{m} [\mathbf{X}_i]_j = L$ for all $1 \leqslant i \leqslant n$. Then the categorization components are defined as normalized entries of these vectors:

$$[\mathbf{f}_i]_k = [\mathbf{X}_i]_k / L,$$

where the components $[\mathbf{f}_i]_k$ are defined in Eq. (1). Let us suppose a learner plays a game with chips i and j, and assigns them to categories v_i and v_j, respectively. If the game is a success, the following operation is performed regarding the component i of the categorization. If $[\mathbf{X}_i]_{v_i} = L$, then nothing changes. Otherwise, the learner updates as follows:

$$[\mathbf{X}_i]_{v_i} \to [\mathbf{X}_i]_{v_i} + 1, \quad [\mathbf{X}_i]_k \to [\mathbf{X}_i]_k - 1,$$

where k is chosen randomly such that $k \neq v_i$ and $[\mathbf{X}_i]_k > 0$. In case of a failure, the categorization of chip i is updated as follows:

$$[\mathbf{X}_i]_{v_i} \to [\mathbf{X}_i]_{v_i} - 1, \quad [\mathbf{X}_i]_k \to [\mathbf{X}_i]_k + 1,$$

where k is chosen randomly such that $k \neq v_i$. Similar operations are performed regarding chip j.

In other words, if a categorization fails, then the value of the category associated with the chip decreases by 1, and another category (chosen at random) is enhanced. In case of a successful game, two outcomes are possible. If the

corresponding category is already full (equal L), then there is no change. Otherwise, the successful category is strengthened (by adding 1); at the same time another (randomly chosen, non-zero) category is reduced by 1.

3.1.4. Reinforcement learner II

A variant of a reinforcement learner has the following update rules: If a categorization fails, then the value of the category associated with the chip decreases by 1 as before, and *all other* categories are enhanced by the amount $1/(m-1)$. In case of a successful game, the successful category is strengthened (unless it is full), and at the same time *all other* non-zero categories are reduced. It turns out that for our purposes, the two types of reinforcement learners behave similarly. Most experiments have been performed by using the first type of reinforcement learners.

Fig. 4 illustrates the concept of a reinforcement learner. In this example, there are only $m = 2$ color categories, such that both variants of the reinforcement learner are the same. Each chip is assigned two non-negative integer numbers that sum to $L = 4$. For instance, chip 1 in Fig. 4(a) has $[\mathbf{X}_1]_1 = 1$ and $[\mathbf{X}_1]_2 = 3$, and chip 2 has $[\mathbf{X}_2]_1 = [\mathbf{X}_2]_2 = 2$. If the first number corresponds to, for example, "light" and the second number to "dark," then with probability $[\mathbf{f}_1]_1 = \frac{1}{4}$ the first chip is categorized as "light", and with probability $[\mathbf{f}_1]_2 = \frac{3}{4}$ as "dark." The second chip is categorized as "light" or "dark" with equal probability $[\mathbf{f}_2]_1 = [\mathbf{f}_2]_2 = \frac{1}{2}$.

Let us suppose that in the first round of the game, chips 1 and 2 are drawn, which are within the similarity range. Let us assume that the learner categorized chip 1 as "dark" and chip 2 as "light." This means that the game failed. The learner will perform the following operations: the "light" stack of chip 1 will go up one reinforcement unit, and its "dark" stack will go down. Similarly, the "light" stack of chip 2 will go down, and its "dark" stack will go up, see Fig. 4(b). In the next round of the game, chips 4 and 26 were chosen; we assume that they are further than k_{sim} apart. The learner assigned categories "dark" and "light" to the two chips, respectively; therefore, this game is a success. The successful update corresponds to strengthening the "dark" stack of chip 4 (and weakening its "light" stack), as well as strengthening the "light" stack of chip 26 (and weakening its "dark" stack), see Fig. 4(c).

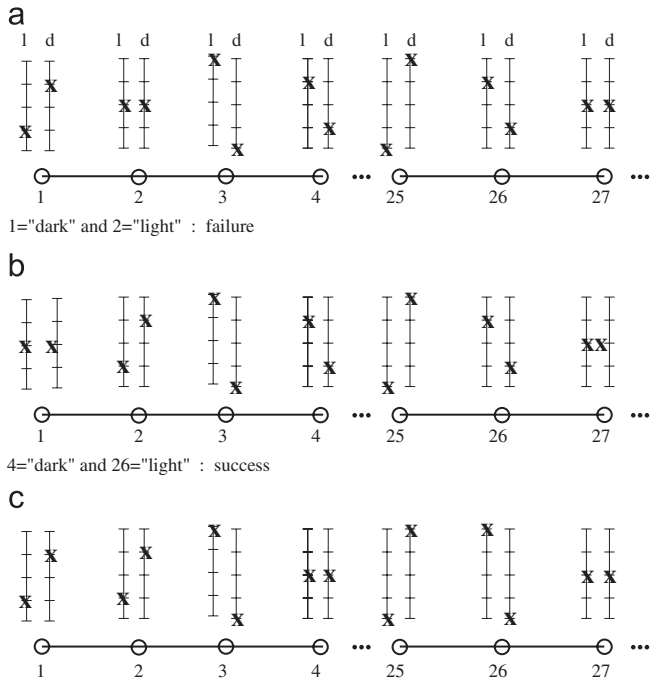

Fig. 4. Two rounds of discrimination–similarity game with a reinforcement learner. Each color chip has two vertical lines or "*stacks*" for categories "light" and "dark" which successively track reinforcement updating. The current values of the stacks are denoted by the **x**. See text for the explanation of the updates performed for the learner.

3.2. Game dynamics and convergence

This subsection investigates the long-term behavior of various color category learners and whether they will, in some sense, learn an optimal system of categorization. Conceptually, the simplest cases for acquiring such categorizations are ones where the color names (signals) and color chips are given in advance and the task is to assign each chip a name so that an effective system of categories emerge, where a "category" is the set of chips signaled by a name. There is obviously an issue as to how the named categories initially become available. But this issue, as well as the one of introducing new names and categories into a system of already existing categories, requires complexities that we do not want to engage at this stage. As mentioned earlier, at this stage it is only necessary to presume that some exogenous or endogenous demands dictate that one or more categories are needed. This is in line with the objective of this article of seeing what results we can obtain about color categorization from considering the simplest kinds of game-theoretic, evolutionary methods.

3.2.1. Discrimination game

Let us fix the initial number of categories, m, and start from an arbitrary color categorization. In particular, we use a random color categorization for deterministic learners (i.e., for the memoryless and smoothing learners). This means that each chip is assigned its category at random through a uniform probability distribution. For non-deterministic, reinforcement learners, the following initial condition is used: each chip has an equal probability to be categorized to each of the m categories. In other words, we initially set for each chip i,

$$[\mathbf{X}_i]_k = L/m \quad \text{where } 1 \leqslant k \leqslant m.$$

We first consider the dynamics of the categorization as the agent plays rounds of the discrimination game. Both linear and circular geometries were used in our computer simulations. We observe the following:

- A memoryless learner does not tend to an optimal categorization (see discussion below).

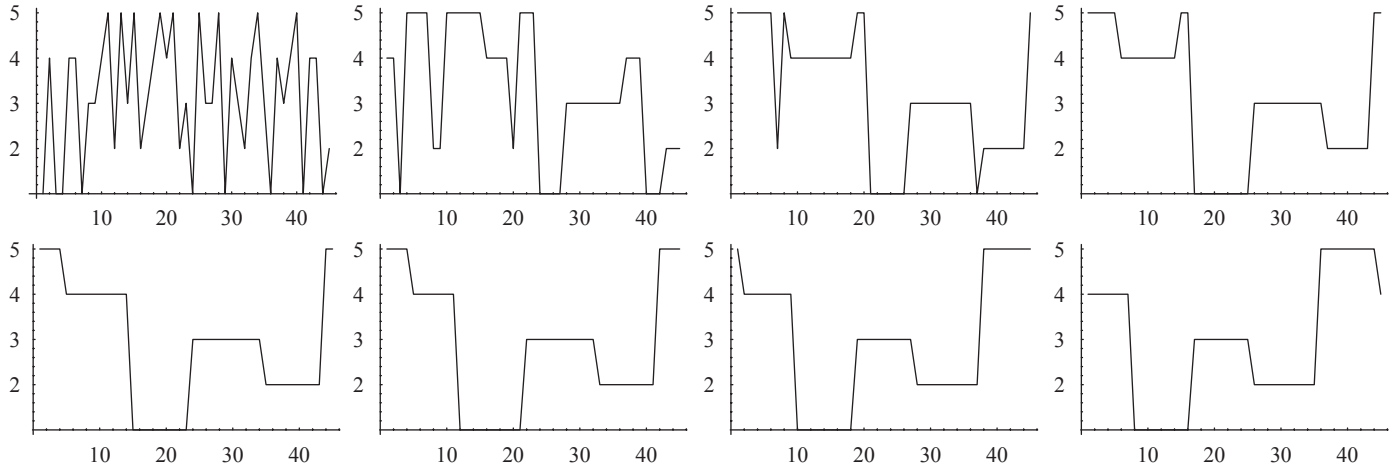

Fig. 5. Several consecutive snapshots of the learning dynamics of a smoothing learner playing the discrimination game, starting from a random initial condition. The horizontal axes represent chips from $n = 1$ to 45, and the graphs show the color assigned to each chip by the current categorization function. We took $m = 5$ categories and $k_{\text{sim}} = 5$. There are 1000 rounds between consecutive snapshots. A leftward drift is noticeable, as seen by tracking the leftmost category (color 5), starting from the top right snapshot. In that snapshot, category 5 is assigned to chips 44, 45 and also chips 1–5. In the next few shots, category 5 moves to the left, reaching range 37–44 in the bottom right snapshot.

- A smoothing learner develops a categorization close to an optimal if $k_{\text{sim}} > 0$. For $k_{\text{sim}} = 0$, as time goes by, a category may be lost, and it can never be regained. Therefore, the only attracting state for $k_{\text{sim}} = 0$ is "one category for all chips"; in practical terms, however, for realistic times color categorizations close to optimal survive (that is, they comprise long-lived states). What is interesting, in the case of a circular geometry, is the resulting categorization is not stationary, and it keeps shifting without changing its color order, see Fig. 5. If the unsuccessful update involves adopting the category from the left (right) neighbor, then the shift proceeds to the right (left). In a randomized case a random drift is observed.
- A reinforcement learner involves the following parameters: the number of chips, n, the number of color categories, m, the similarity range, k_{sim}, and parameter, L, the sum of the components of the vectors, $\mathbf{X}_i$. Depending on these parameters, different convergence behaviors are observed. For $m = 2$, the learner always gives rise to a steady (non-shifting) categorization close to an optimal one. The convergence rates depend on the parameters such that they are faster for larger k_{sim}. For low values of k_{sim}, L must be large to achieve convergence. For $m > 2$, the algorithm does not converge in a reasonable time.

Note that throughout this article we will informally use terms such as "realistic time," "close to optimal," etc. They have a clear intuitive meaning but we do not attempt to quantify them at this stage of model development.

3.2.2. Discrimination–similarity game

As shown above, the discrimination game leads to a close-to-optimal categorization only for smoothing learners. The resulting categorization is not stationary, but it drifts in one or both directions. The next set of experiments examines the dynamics of convergence for discrimination–similarity games. Again, linear and circular geometries were used. We observe the following behavior:

- A memoryless learner does not tend to an optimal categorization.
- A smoothing learner develops a "drifting" categorization close to an optimal, as before.
- A reinforcement learner tends to a stationary (without a directional shift) categorization which is close to an optimal categorization, see Fig. 6. There is no convergence in the strict sense, because the dynamics does not have any fixed points. However, the solution remains in the vicinity of a nearly-optimal configuration for a long time.

3.2.3. The failure of the memoryless learner

Historically, the memoryless learner algorithm has been used as a very simple algorithm which makes minimal demands on the learner's "cognitive" apparatus but nonetheless achieves its learning goal (converges) in many settings. That is why we considered this algorithm as a null-hypothesis of our argument. It was therefore somewhat of a surprise that in this study, the memoryless learner does not find an optimal color categorization in either of the games.

This is different from the behavior of the memoryless learner algorithm when it is applied to learning fixed categories, which is the usual setting for its implementation in artificial intelligence research (Niyogi, 1998). There, one envisages a series of interactions between a learner and a teacher. There is a (finite) number of concepts (rules, grammars, etc.) and the teacher knows the "correct" one.

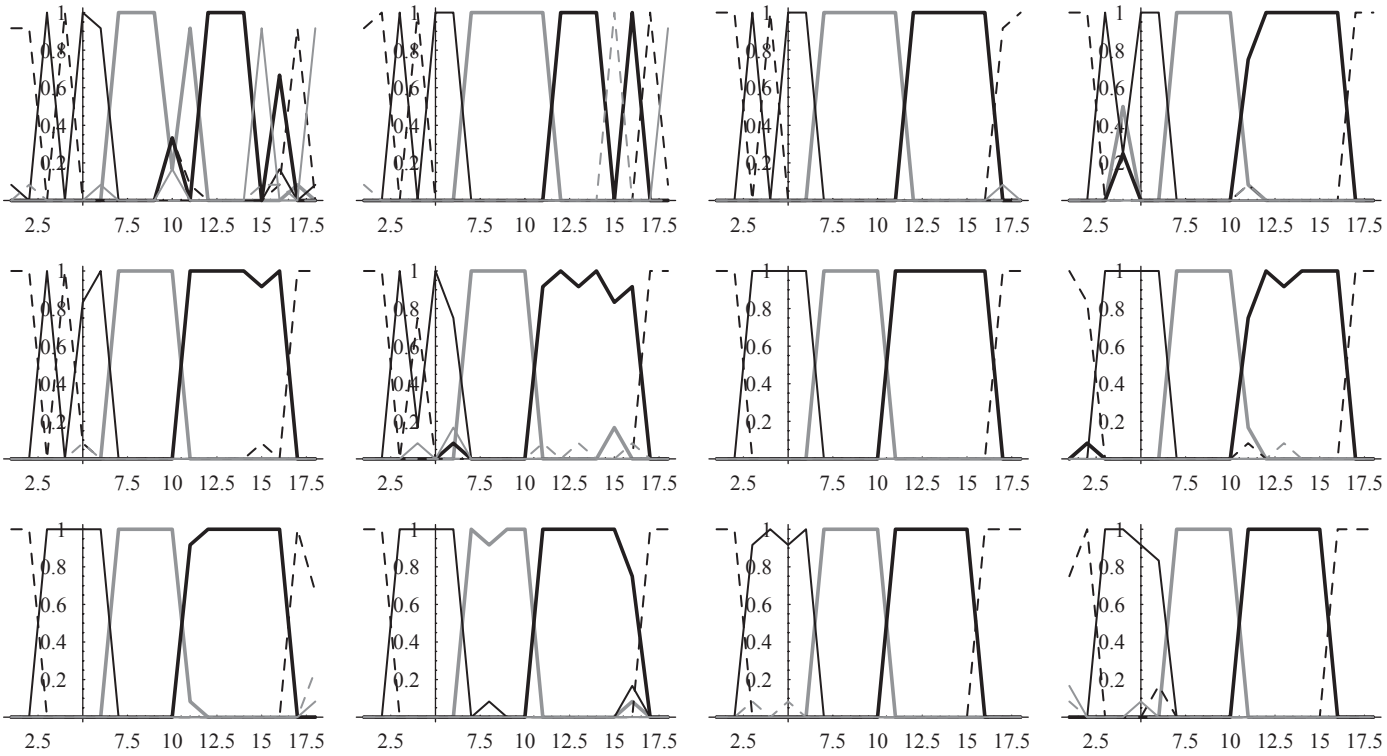

Fig. 6. Several consecutive snapshots of the learning dynamics of a reinforcement learner playing the discrimination–similarity game. The horizontal axes represent $n = 20$ chips, and the vertical axes are the probability for each of the color categories to be chosen for each chip. We took $m = 6$ color categories, $L = 12$ and $k_{\text{sim}} = 3$. The initial condition is such that each of the 6 color categories has an equal probability to be chosen for each chip (not shown). The thick, thin and dashed, both black and gray, lines correspond to the 6 color categories. There are 4000 rounds between consecutive snapshots.

The learner's task is to guess the correct concept by evaluating a string of examples (applications of the rule, grammatically correct sentences, etc.) given by the teacher. The learner starts from a randomly chosen first guess and receives the first input from the teacher. If this is consistent with the learner's hypothesis, no action is taken. If it is inconsistent, the learner adopts a different, randomly chosen hypothesis. One can prove that (under some mild conditions on the underlying set of hypotheses) as the number of such games goes to infinity, the memoryless learner will converge to the right answer (Komarova & Rivin, 2003).

In order to appreciate similarities and differences in learning between the above teacher–learner paradigm and our memoryless learner paradigm, consider our memoryless learner as a set of n agents, each trying to learn the correct category for its chip. The difference is that in our case there is no fixed "correct" categorization, and, as the number of games tends to infinity, there will always be failed games, no matter what the current categorization is. Indeed, even for the optimal solution, some of the learners will inevitably find themselves on the boundaries of color domains, and for them many discrimination–similarity games will fail, simply because the two neighboring colors will be assigned to different categories. This situation is reminiscent of a conventional memoryless learner trying to learn from an inconsistent teacher, who gives contradictory cues of what the correct answer might be. A memoryless learner is notoriously unsuccessful in such settings (Niyogi, 2006), which is consistent with our result.

3.2.4. Gaining and losing color categories

We have observed that the estimate of Eq. (8) holds true, and the learning dynamics of a reinforcement learner may lead to color category emergence or color category extinction. If we start from the number of categories, $m > m_c$, we observe that some categories are wiped off (or at least are driven to very low probabilities), and the total number of active color categories corresponds to m_c. In Fig. 6, we start from $m = 6$ colors while $m_c \approx 4.1$. We observe that most of the time, there are only four dominant color categories present. The other two categories exist at low probabilities, and sometimes they come up at the color category "junctions" as shown in Fig. 6. Similarly, we could start from $m < m_c$, add a small chance to use an additional color category and observe an increase in the number of dominant color categories, Fig. 7. Adding an additional color category starting from $m > m_c$ does not lead to an increase in the number of categories (not shown); on the contrary, m decreases to reach m_c.

3.2.5. A note on the speed of convergence

In this article the convergence rates for various algorithms are not calculated, because the main goal here

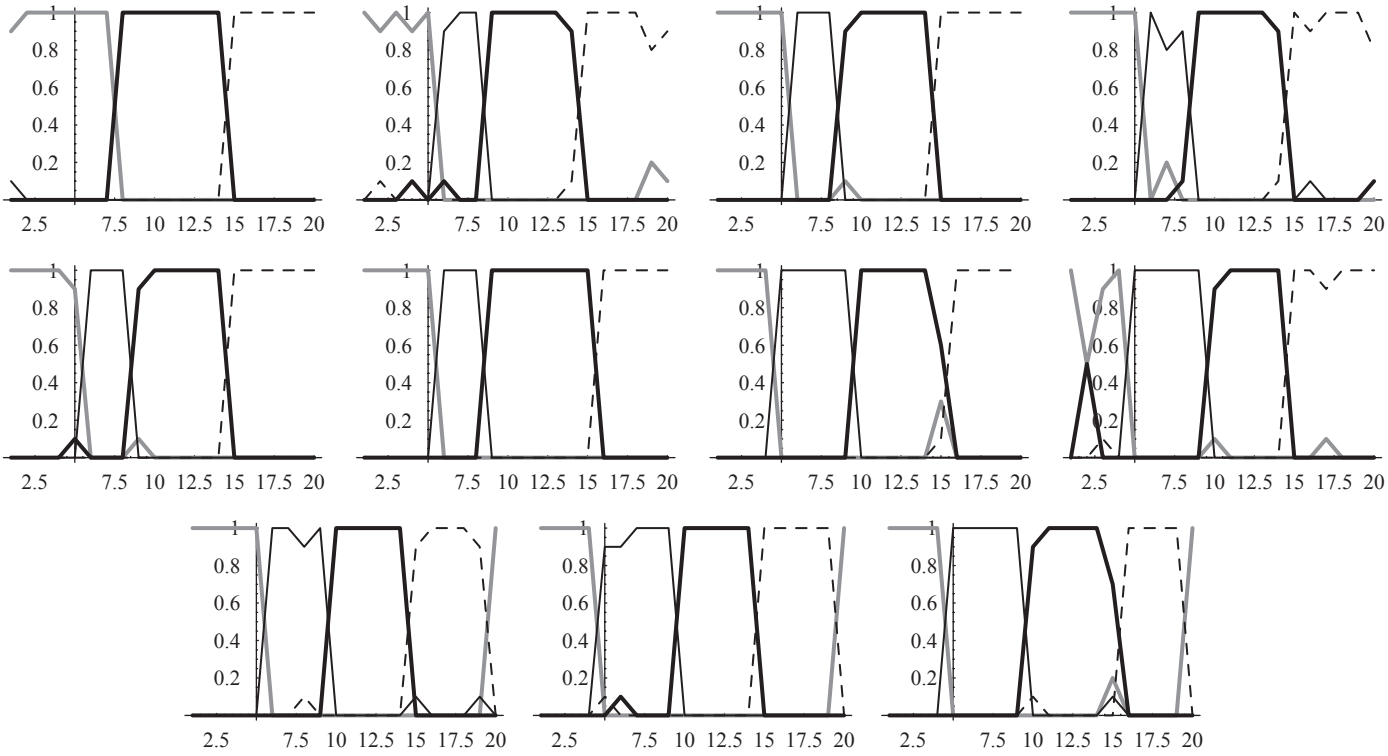

Fig. 7. Several consecutive snapshots of the learning dynamics of a reinforcement learner playing the discrimination–similarity game. Here, we took $L = 12$ and $k_{\text{sim}} = 3$. Initially, we have three dominant, deterministically assigned color categories, with a small perturbation introducing a small probability of the fourth color category for the first chip. The thick black, thick gray, thin black, and dashed lines correspond to the 4 color categories. There are 4000 rounds between consecutive snapshots.

is to demonstrate that certain kinds of individual learning algorithms produce nearly optimal solutions. However, it is worthwhile to comment about the time it takes to weed out extra categories as both n and the optimal number of categories increase. Fig. 8 shows a nearly optimal categorization reached by a reinforcement learner playing rounds of the discrimination–similarity game, starting with $m = 8$ color categories. It took on the order of 10^6 rounds to settle to approximately 6 categories. This should be compared with the dynamics of Fig. 6, where the learner went from 6 to the optimal 4 color categories after about 10^4 rounds.

3.3. Inhomogeneous color diet and variable color importance

So far we assumed that the color chips are drawn from a uniform distribution. Color scientists talk about different "color stimulus diets" in which certain parts of the color space are more frequent or more salient than others (e.g., Regan et al., 1998; Stoner, Riba-Hernández, & Lucas, 2005). Analogous to this, some of our simulations involve non-uniform distributions of color chips. Namely, we assumed that the color distribution included a region of frequently observed colors, or a color "hot spot." For example, in the experiments of Fig. 9, in 50% of the cases, a color chip was drawn from a "hot spot" region in the color stimulus domain (a continuous region of 10 chips out

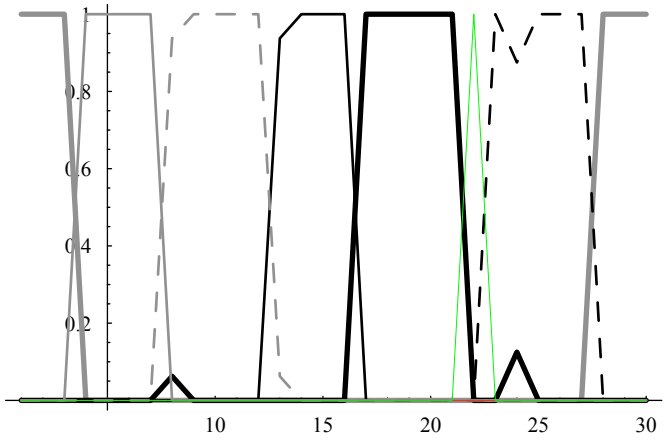

Fig. 8. A snapshot of the reinforcement dynamics of discrimination–similarity game, with $n = 30$ chips, $L = 16$, $k_{\text{sim}} = 3$, and 8 color categories equally distributed initially. $m_e = 6.1$ in this case. This shot corresponds to about 10^6 iterations of the game.

of the total of $n = 40$ chips). In the remaining 50% of the cases, a chip was drawn from a uniform distribution over the whole color domain. In Fig. 9 we observe that the first categories that emerge are the ones that surround the "hot spot." Eventually, $m = 6$ color categories will clearly develop (not shown).

The "hot spot" simulation just described emphasizes the effect on categorization of visiting a portion of the color

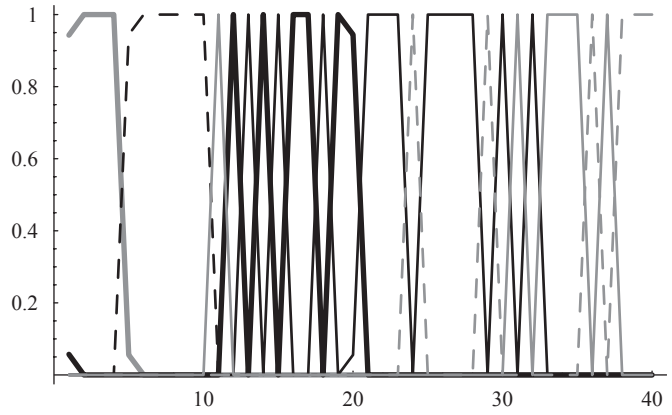

Fig. 9. Discrimination–similarity game of a reinforcement learner with a non-uniform color distribution. The leftmost 10 chips (the "hot spot") are chosen in 50% of the cases (see text). Parameters are $n = 40$, $m = 6$, $L = 15$, $k_{\text{sim}} = 4$. The frame is taken after about 10^5 runs.

space more often. By comparison, the simulation described below investigates the effect on categorization of allowing k_{sim} to pragmatically establish importance in a portion of the color space.

In all the above mentioned investigations we used the simplifying assumption that k_{sim} is a constant for all color chips. This has been interpreted as follows: for pragmatic purposes, colors that are within the k_{sim} range will most likely have similar pragmatic importance. However, in the primate world the importance of color similarity may be different in different regions of the color space. For example, it may be of practical importance to distinguish shades of colors in the red–yellow range, and less important to notice differences in the blue–purple region. The reason for this could be related to the kinds of edible fruit available. If valued edible fruit are reddish, and almost nothing edible (or dangerous) is bluish, then the reddish region becomes more important, and subtler differences are adaptively acquired for that region. Whatever the reasons, the literature contains considerable empirical evidence to support the present suggestion regarding non-uniform color salience in non-human primates (Stoner et al., 2005; Regan et al., 1998) including observed perceptual non-uniformities in human color appearance space (e.g., Kuehni, 2004; Malkoc, Kay, & Webster, 2005).

The next set of experiments investigates such situations. We draw chips from a uniform distribution over $n = 40$. If at least one of the chips in a pair belongs to the range 1–10, then we set the similarity range to be $k_{\text{sim}} = 2$. Otherwise, k_{sim} is taken to be 6, see Fig. 10.

The simulation starts with $m = 6$ color categories, each equally likely to be chosen for every color chip. After about 10^5 runs, the following picture starts to emerge: the region 1–11 is divided into 3 color categories, and there are also a couple of larger color categories that correspond to the other 29 chips. What we observe is a non-homogeneous color categorization with finer categories in the region of small k_{sim} and rougher categories in the rest of the field. This result is quite predictable. We can calculate the optimal number of color categories for the 10 chips with $k_{\text{sim}} = 2$, $m_c = 2.9$, and the optimal number of categories in the rest of the circle, with $n = 30$ chips and $k_{\text{sim}} = 6$: $m_c = 3.3$. Thus we expect to have roughly three categories for chips 1–10 and three categories for chips 11–40, see the last frame in Fig. 10.

All simulations up to this point have dealt only with *individual learning*, most directly capturing categorization within a single artificial agent or some empirical situations involving the learning of color categories without communication. To model the evolution of *shared* human-like categorization systems used for pragmatic color communications, we next consider category solutions that emerge when individual agents in a *population* communicate in color category games.

4. Population learning

This section looks at the evolution of color naming in a population of agents. The agents play versions of the discrimination and the discrimination–similarity game through interactions with one another. Interestingly, we observe that learning algorithms that did not find an optimal categorization in an individual learning task, do not in general improve their performance in a population-learning setting. One might conjecture that interactions among individuals could improve convergence properties of learning algorithms. This did not happen: for instance, a population of memoryless learners is unsuccessful in developing an optimal color categorization. The following two subsections consider smoothing learners playing rounds of discrimination game and reinforcement learners playing rounds of discrimination–similarity game.

4.1. A population of smoothing learners playing a discrimination game

Let us suppose we have N individuals, $\{1, \ldots, N\}$ in the population, each having its own deterministic categorization, $F^{(I)}$, $1 \leqslant I \leqslant N$. Time flows discretely. At each time-step, two individuals from the population are picked at random and presented with two color chips chosen at random. First, each individual plays a discrimination game using the two chips. There can be three outcomes of the games, which define the update rule of the player:

(1) One of the individuals succeeds and the other fails. In this case the failing individual learns the color categories (for the two chips) from the former individual; that is, the successful individual is taken as the teacher and the failing one as the learner.
(2) Both individuals succeed in discriminating the two chips. In this case one of the two individuals is picked at random to be the teacher. The other one learns the color categories of the two chips from the teacher.
(3) If both individuals fail, then both of them perform the unsuccessful update from the smoothing learner

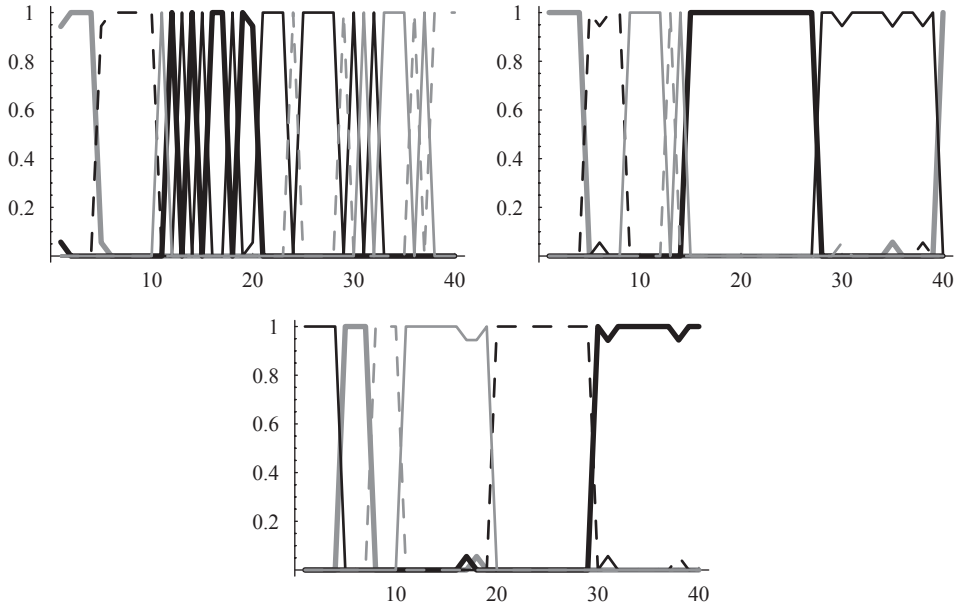

Fig. 10. Discrimination–similarity game of a reinforcement learner with a non-constant k_{sim}. Parameters are $n = 40$, $m = 6$, $L = 15$, k_{sim} varies between 2 for chips 1–10 and 6 for chips 11–40. The last frame is taken after about 10^7 runs.

algorithm. (As a variant of this algorithm, only one of the failing individuals could perform the unsuccessful update procedure.)

An example of a simulation for a population of smoothing learners is presented in Fig. 11. There, each row of plots corresponds to one learner. The first column represents 10 individual population solutions at the initial time, and the second and third columns show snapshots of solutions at two later moments of time. For each color chip i (the horizontal axes), only a single color category is shown, which corresponds to the maximum entry in the vector $\mathbf{X}_i$. In other words, for each player and for each chip, we show the color category that the player is most likely to assign to it. For instance, player 1 (shown as the first row of plots in Fig. 11) is most likely to use color category 3 for chips 1–11 after 10,000 rounds, whereas he is most likely to use category 2 for the same chips after 30,000 rounds (as seen in the second and the third columns).

We can see that the smoothing learner algorithm converges to a high-coherence population with a color categorization close to an optimal one. By *high coherence* we mean a high degree of agreement among players on their choice of color categorization. If the smoothing algorithm implies adopting the color category from the neighboring chip on the right (left), then a slow synchronized leftward (rightward) drift of color categorizations is observed in the entire population.

A variant of this algorithm includes the notion of fitness. Here, we define the fitness of individual I as

$$\phi_I = \frac{1}{Nn}\sum_{J=1}^{N}\sum_{i=1}^{n}\delta_{F_i^{(I)},F_i^{(J)}}, \tag{9}$$

where δ stands for the Kronecker symbol. This definition gives the degree to which an individual agrees with others regarding color categorization. Note $0 \leqslant \phi_I \leqslant 1$. Now, each time both players succeed in the discrimination game, the one with the larger fitness is chosen as a teacher; if the fitnesses are the same, then the teacher is randomly chosen. We have run a series of experiments where we included fitness in the dynamics. No qualitative difference in the results has been observed.

A problem with the discrimination game is that the population can learn an arbitrary number of categories $m \leqslant n$. That is, an optimal solution can be "each chip has its own category." In fact, if we play the discrimination game with different given numbers of categories, m, then the success rate of the games with $m = n$ will be the highest. This is an outcome that is not compatible with the goal of modeling properties of human categorization behavior, because it suggests the unrealistic outcome that different color names for all distinguishable color chips will evolve.

4.2. A population of reinforcement learners playing a discrimination–similarity game

Suppose there are N reinforcement learners, each equipped with n vectors, $\mathbf{X}_i$. As before, a pair of individuals is chosen from the population. They are presented with two chips chosen at random. There are three cases:

(i) The two chips are the same.
(ii) The two chips are different but within k_{sim} of each other.
(iii) The two chips are different but not within k_{sim} of each other.

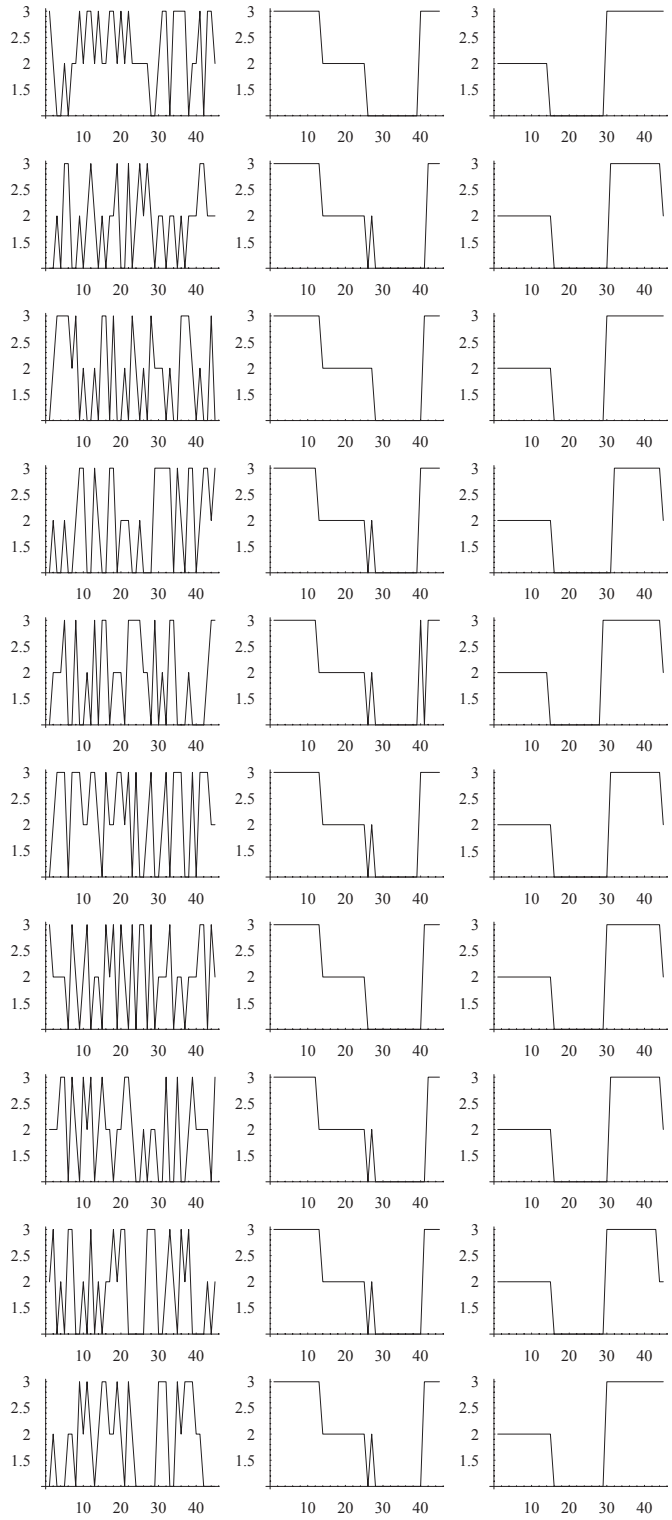

Fig. 11. A population of smoothing learners playing the discrimination game, with $N = 10$ agents, $n = 45$ color chips, $m = 3$ and $k_{\text{sim}} = 5$. The first column is the initial, random, color categorization of the 10 players. The second column is the categorization of the players after 10,000 rounds of the game. The last column is the categorization 30,000 rounds later. For each chip (horizontal axes) only one color category is shown: the one that the individual is most likely to use for this chip.

Table 1
Four possible outcomes of each round of the discrimination–similarity game

	Same chip	Chips within k_{sim}	Chips further than k_{sim}
Same–same	C	C	D
Same–different	N/A	B	A
Different–same	N/A	A	B
Different–different	N/A	D	C

The left column lists the categorization choices of the two players. A denotes that player 1 fails and player 2 is successful, B denotes that player 2 fails and player 1 is successful, C denotes that both are successful, and D denotes that both fail.

Each player assigns categories to the two chips. Depending on the players' choice of categories, their game is a success or a failure. For instance, assigning the same category to the two chips in the case where they are within the k_{sim} range comprises a success; and if the chips are further apart than k_{sim}, it is a failure. None, one, or both players can be successful in any given round of the game. There are four outcomes which we denote A, B, C and D; they are summarized in Table 1. Each of the four outcomes has its own update rule:

(1) In the case of outcome A, player 1 learns from player 2. This indicates that the vector of player 1 corresponding to the first chip (chip i) changes according to this rule:

$$[\mathbf{X}_i]_{v_1^{(2)}} \to [\mathbf{X}_i]_{v_1^{(2)}} + 1, \quad [\mathbf{X}_i]_{v_1^{(1)}} \to [\mathbf{X}_i]_{v_1^{(1)}} - 1,$$

where $v_1^{(1)}$ is the category chosen by player 1 for chip i and $v_1^{(2)}$ is the category chosen by player 2 for the same chip. In other words, player 1 (the learner) updates his categorization for the first chip in the following way: he adds 1 to the stack corresponding to the color category used by player 2 (the teacher), and he subtracts 1 from the stack corresponding to the color he previously used for this chip.

Similarly, player 1 updates the component corresponding to the second chip:

$$[\mathbf{X}_j]_{v_2^{(2)}} \to [\mathbf{X}_j]_{v_2^{(2)}} + 1, \quad [\mathbf{X}_j]_{v_2^{(1)}} \to [\mathbf{X}_j]_{v_2^{(1)}} - 1.$$

(2) In the case of outcome B, the roles are reversed, and player 2 learns from player 1.
(3) In the case of outcome C, in the simplest algorithm, the teacher is chosen at random.
(4) In the case D, each of the players update their vectors according to the individual reinforcement learner algorithm.

Note that the task of learning color categories of one player from another has to be modeled differently for smoothing and reinforcement learners. Smoothing learners are deterministic learners, that is, they assign a particular color category to each chip. Therefore, the only learning

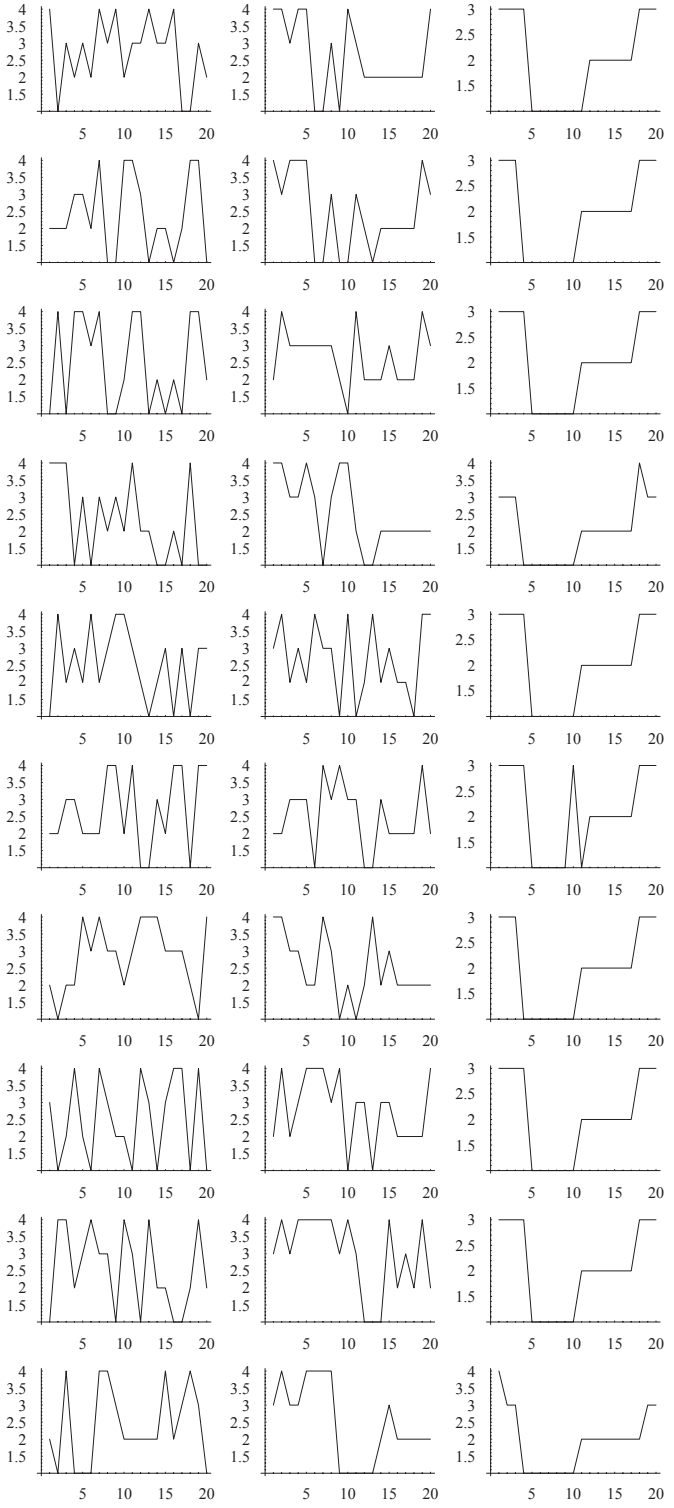

Fig. 12. A population of reinforcement learners playing the discrimination–similarity game, with $N = 10$ agents, $n = 20$ color chips, $m = 4$, $L = 12$ and $k_{\text{sim}} = 4$. Each row of plots corresponds to the color categorization of one player. For each chip (the horizontal axes) we show the category which has the highest entry in the learner's categorization vector. The first column is the initial, random, color categorization solutions of the 10 players. The second column shows the categorization solutions of the players after 10,000 rounds of the game. The last column shows the categorizations 70,000 rounds later.

mechanism available for them is to switch the color category for the given chip to that of the teacher. On the other hand, reinforcement learners allow for a more realistic modeling of the learning process. Unlike smoothing learners, they use probabilistic categorization. As the result of learning, the failing individual increases the "stack" corresponding to the color category used by the teacher. Thus the next time this chip is chosen, this individual will have an increased chance to assign it to the teacher's category.

This reinforcement learner algorithm gives rise to a coherent population of stationary optimal categorizations, see Fig. 12. In the figure, each row of plots corresponds to one learner. As in Fig. 11, for each color chip i (the horizontal axes), only the most likely color category is shown. As column three shows, we have $m_c \approx 3$ in this case, that is, only 3 color categories survive.

Note that in the algorithm described above, we included pairs of identical chips, see Table 1. Such pairs were always discarded in the previous sections in which individual learning was discussed. In an additional series of experiments, we discarded pairs of identical chips in population learning scenarios. This did not affect the results in a qualitative way.

We have considered two other extensions of this algorithm, both concerning outcome C, where both players are successful in the game. First of all, each player can keep a score, that is, count the number of times he served as a teacher. Then, when it comes to a tie (C), the player with a higher score (rather than a random player) is chosen as a teacher. (If both players have the same score, then the teacher is chosen at random.)

The other extension uses the notion of fitness of the player. In case C, the player with a higher fitness is chosen as a teacher. The definition of fitness for non-deterministic agents is given as follows. We first define the quantities

$$F_i^{(I)} = \kappa, \quad [\mathbf{X}_i^{(I)}]_\kappa = \max_k [\mathbf{X}_i^{(I)}]_k,$$

that is, for each color chip i, we pick the category $\kappa \in \{1, \ldots, m\}$ such that the component κ of the vector $\mathbf{X}_i^{(I)}$ is the largest. In the case where several components have the same magnitude, we could pick one at random. Then the fitness can be defined by Eq. (9).

It turns out that these extensions of the algorithm do not produce a qualitative difference on the outcome of the evolutionary dynamics compared to those that do not use *fitness* or a *success score*.

5. Discussion

5.1. Possible extensions of the evolutionary algorithms

Our approach in this article has been to consider some very simple idealizations of situations where a signaling system evolves for objects from a continuous domain. This subsection discusses some of the ways the concepts and

algorithms of the previous sections can be modified so that they apply to more complicated situations. The modifications are only briefly discussed, and details involving exactly how the evolutionary algorithms are to be extended are not presented. They will be developed in future publications.

5.1.1. Extending algorithms for circle and line stimulus domains to a three-dimensional color solid

As detailed above, the present investigations only consider categories formed on continuous circle and line segment gradients. It is straightforward, however, to extend the present investigations to simulating category evolution on a three-dimensional color solid, like the Munsell color solid, with minor modifications of the algorithms.

That is, chips a and b are said to be (k,l,m)-*similar* if and only if simultaneously, (i) in terms of the hue dimension a and b are k-similar, (ii) in terms of the value dimension a and b are l-similar, and (iii) in terms of the chroma dimension a and b are m-similar. By appropriate substitutions of (k,l,m)-similarity for k-similarity in our algorithms and other concepts, our evolutionary methods of analyses extend to various two- and three-dimensional regions of the Munsell color solid.

5.1.2. How agents can immediately achieve correct categorization of new stimuli

It is reasonable to consider cases where the number of stimuli is so huge and diverse with respect to the number of signals that agents experience only a small fraction of the possible stimuli. Our algorithms, as currently formulated, do not apply to such cases, because they require each chip to be updated, usually a large number of times. One approach to extending the algorithms to cover these cases is to evolve for each name an *icon chip*. Intuitively an icon chip approximates one feature of human long-term memory in the naming of a newly presented color chip. Formally, any chip c occurring in a game is named as follows:

(1) If there is no icon chip to which c is k-similar, play the game, give c the name dictated by the result of the game, compute the categorizing vector $\mathbf{f_c}$ accordingly, and make c an icon chip.
(2) If i is the only icon chip to which c is k-similar, play the game for c with c having i's name, recompute $\mathbf{f_i}$ to reflect the result of the success or failure of the game (even though the game was played with c), remove i as an icon chip and replace it with a new icon chip i' that is a chip that is nearby i in the direction of c, and set

$$\mathbf{f_{i'}} = \text{ the recomputation } \mathbf{f_i}.$$

(3) If there are more than one icon chip to which c is k-similar, randomly select one of those icon chips, i, play the game for c with c having i's name, recompute $\mathbf{f_i}$ to reflect the result of the success or failure of the game (even though the game was played with c), and retain i as an icon chip.

In this approach, chips that are judged are only given names of icon chips, and only icon chips are updated. The approach may be viewed as a means of incorporating a primitive form of perceptual memory into the evolutionary process, with the presented chips being analogous to "perceptions" and the icon chips to "memories."

5.1.3. Best exemplars

The introduction of icons as just described allows for the possibility of evolving best exemplars for color categories. In such a situation, an icon chip is an obvious choice for a best exemplar. However, there are likely to be several icon chips for a given color category. When a color-naming system achieves a near equilibrium state in a discrimination-similarity game using k-similarity, the icon chips nearest a category's boundary are at a disadvantage for being a best exemplar for that category, because they may sometimes name a chip that is classified as a failure in simulated games. Exactly which of a category's icon chips should be selected as "best exemplar" would generally depend on additional factors not emphasized in the simulations presented in Sections 2 and 3, for example, heterogeneity of chips, heterogeneity of agents, variable k-similarity, etc.

5.2. Implications for color-naming theory

It is important to re-emphasize that the agent simulations we present are not intended to model human color category learning or interactions between human categorizers. They are instead intended to demonstrate what can be achieved using only the most rudimentary forms of color observation and communication together with an elementary evolutionary dynamics. The evolutionary dynamics used follow the ideas that (i) color naming should be based on pragmatic concerns, (ii) in general, perceptually similar colors should be given the same name, and (iii) perceptually different colors should be given different names. These kind of simulations may be useful for clarifying certain contentious issues in the literature concerning the basis for color naming. This may be done by providing counter-examples which show that various features of naming systems can evolve without making additional assumptions involving physiological processing, cognitive strategies, or socio-cultural methods of transmission. On the positive side, by having explicit evolutionary models and algorithms we may be able to demonstrate the feasibility of certain evolutionary theories presented in the literature.

Another goal of the present work was to begin investigations into some predictions made by the IDM of color categorization (Jameson, 2005d; Jameson & D'Andrade, 1997), and to evaluate how such predictions hold up

under simulated situations of category evolution. We believe that beyond the limited assessment possible using diachronic analyses of color lexicon evolution, the approach presented here is perhaps one of the few ways to evaluate theories of color category evolution, since directly observing or assessing the evolution of such systems in the real world is not achievable via typical psychological or cultural investigative methods. Still, despite our admittedly indirect evaluation of the IDM, we found several fruitful results that bear both on color-naming psychology and extensions of some of the information processing principles described by Garner (1974).

First, our discrimination-similarity games with uniform sampling and reinforcement dynamics produced equal sized category partitions. We view this as support for the prediction that successful color-naming systems exhibit "…an informational advantage to making the divisions so that category foci are maximally different from each other." (Jameson & D'Andrade, 1997, p. 313). That is, although here category partitions arose in the absence of defined category "best-exemplars," the stable relational structures seen among our categories implies that the best-exemplars of a given category partition are not close to neighboring category boundaries (a feature dependent on k-similarity). This finding is also compatible with categorization dynamics described by Garner (1974).

In addition, investigations that explored the impact of inhomogeneous color space sampling and inhomogeneous k-similarity ranges on individual agent categorization validated the *IDM* prediction that exogenous pragmatic influences (such as *hotspots*) affect individual's category partitioning in ways that can trump other psychologically based tendencies that shape categorization.

5.2.1. Category solutions under inhomogeneous color space sampling and inhomogeneous agent similarity ranges

In general, under conditions of uniformly distributed color chips we find that the convergent solutions produce equal sized categories (where "size" is measured in terms of jnds). Predictions regarding a tendency toward equal sized category regions were presented in the *IDM* (Jameson, 2005d, p. 320). In addition, however, we investigated two cases illustrating factors that influence the emergence of category regions of equal size.

For example, Section 3.3 presents algorithmic solutions under (i) varying distributions of color chips and (ii) varying color similarity range parameter settings. For situation (i) some segments of the hue continuum were sampled more liberally than other segments when setting up the games to be played, following the rationale that pragmatic concerns (i.e., colors signaling ripe fruit) might present sampling biases that could impact category development. In situation (ii), an agent's similarity range parameter was inhomogeneous across the color continuum.

Two interesting results emerged from situations (i) and (ii): First, the segments of the hue continuum in which more games were drawn were found to form categories earliest. Second, the number of categories found in the region of smaller k_{sim} was larger, allowing the agents to use finer distinction among color shades in that region.

The first result may provide a hint regarding the unexplained widespread occurrence of the early emergence of reddish categories in human color categorization systems. By analogy, if a pragmatic concern of optimizing caloric intake is an especially important factor in a population's color signaling system, then the categories most salient for this concern (e.g., colors for ripe fruit) may emerge first and stabilize earliest.

The second result suggests that when differences in agents' similarity ranges exist, the convergent categorization solution can shrink both the size and the number of categories to a system that is near optimal.

Both these results are important for evaluating how pragmatic constraints on color naming might influence the evolution and maintenance of a color signaling system in both artificial agents and humans. They also indirectly give an impression how systematic variation in observer-type heterogeneity could influence convergent category solutions (as suggested by Jameson, 2005a)—a topic of recent discussion in the color categorization literature (Steels & Belpaeme, 2005). Finally, both results accord with the organizational framework for human color categorization described by Jameson (2005d, pp. 316–325).

5.2.2. Relevance to the color category simulation literature

The modeling methods used here resemble those found in existing research on the evolution of general communication and signaling systems (Grice, 1957, 1989; Komarova, 2004; Komarova & Niyogi, 2004; Komarova, Niyogi, & Nowak, 2001; Lewis, 1969; Nowak, Komarova, & Niyogi, 2001; Nowak et al., 2002; Skyrms, 1996) and from computational modeling specific to perceptually based color categorization (Belpaeme & Bleys, 2005; Steels & Belpaeme, 2005; Zuidema & Westermann, 2003).

Similar to the color category simulation research (Belpaeme & Bleys, 2005; Steels & Belpaeme, 2005), we examine how individual agents behave in color discrimination games and how groups of agents interact in language games under differing constraints. Our communication games allow a shared lexicon to emerge and stabilize across simulated communication games among a population of agents.

In particular, Steels and Belpaeme (2005) recently investigated the circumstances under which simulated agents could arrive at human-like color categorization solutions. Their goal was to explore the potential for agent communication with humans.

They implemented algorithms that incorporated the standardized three-dimensional model of human color perception (i.e., *CIELAB*). They investigated (i) whether these algorithms evolved category systems that were sufficiently shared among agents to allow successful communication in the simulated population, and (ii)

whether the evolved category systems resembled those systems found for humans.

By comparison, the definition of an agent in our investigation differs from that of by Steels and Belpaeme (2005) in that we do not incorporate a large and rich set of human color perceptual features into our simulated agents; instead our agents incorporate only a primitive ability to carry out color discriminations.

Interestingly, Steels and Belpaeme (2005) conclude that the collective choice of a shared repertoire must integrate multiple constraints, including constraints coming from communication. One can argue that the present simulations actually make this point more forcefully, because we found stable shared categorization systems that only required simple, basic pragmatic assumptions about communication simple, basic learning rules, and only a rudimentary assumption concerning an agent's discrimination abilities.

In addition, we differ from Steels and Belpaeme (2005) regarding the specification of the stimulus set evaluated by agents, because, unlike them, we do not use complex information about real-world surface reflectances as the input stimulus sets in our simulations. Instead we limit the stimulus domain to simple stimuli organized in terms of jnds. The present study employs only stimuli arranged in circle and line gradients. However, as suggested previously, we expect our methods will extend in simple ways to the full Munsell solid. The kind of constraints we employ do not aim to capture the considerably more complex sets of constraints occurring in real-world color categorization among populations of humans. This latter point does not diminish the significance of our strategy for understanding the real-world color naming in terms of evolutionary models involving minimal perceptual constraints and simple learning. The rationale behind investigating the evolution of color categories this way is that if convergent stable color categories are not observed under such evolutionary processes, then we know that additional constraints (e.g., actual distributions of environmental colors, a simulated model of human color perception, more realistic language learning processes, etc.) or more complex learning algorithms are needed for determining the conditions under which a stable, convergent system will emerge. The outcome of our simulations, however, found that such additional constraints were not required to produce category learning and stable convergent categorization systems.

5.3. Implications for social evolution

Signaling learning games within a population have been used by philosophers (e.g., Lewis, 1969; Skyrms, 1996) as devices for illustrating, arguing for, and refuting various positions about the nature of social evolution. However, in formulating their arguments they did not consider games that converged to a near optimal solution. The existence of such convergent behavior—for example, in our population learning game involving the color circle with smoothing learners—reveals shortcomings of some important definitions employed in the literature and presents some new and interesting avenues for modeling.

The philosopher Lewis (1969) was the first to use signaling games to illustrate and formulate social evolutionary concepts. He provided a game theoretic account for the formation of conventions and used it as the basis for a theory of meaning for signaling systems. Skyrms (1996) later provided a more penetrating analysis of the evolution of meaning within signaling systems. Both base their theories on the game theoretic concept of "equilibrium," or more precisely, "a robust Nash equilibrium of a coordination game." Lewis in his final definition of "convention" (he used 73 pages of text to arrive at this final definition) allows for the equilibrium to almost hold instead of exactly holding. In the following quotation from Lewis (1969), P stands for a population of agents and S for a convention:

> There is no harm in allowing a few abnormal instances of S which violate some or all of the clauses [of Lewis's definition of convention as a Nash equilibrium of a coordination game]. So we replace "in any instance of S of members of P" by "in almost any instance of S among members of P." If we want more precision, we can replace it by "in a fraction of at least d_0 of all instances of S among members of P" with d_0 set slightly below one.
>
> Nor is there any harm in allowing some, even most, normal instances of S to contain a few abnormal agents who may [violate the conditions of S being a convention]. (p. 77)

Having a convention to hold almost universally rather than universally does not necessarily lead to an "almost equilibrium of a coordination game," which in our reading of Lewis is the intended conclusion to be drawn, but could possibly lead to conventions that do not stay close to any particular solution. We observed both kinds of behaviors in our population learning simulations: For a population of reinforcement learners playing a discrimination–similarity game, we observed a solution that remains in the vicinity of a nearly optimal configuration for a long time; for a population of smoothing learners, we observed a non-stationary convergence to a near optimal categorization that slowly drifts outside of the vicinity of that categorization and into the vicinity of another optimal categorization.

In the population of smoothing learners, unsuccessful updates occurred either (i) through a random choice, producing random drifts, or (ii) choosing the left (right) neighboring category, producing shifts to the right (left). In an evolutionary scenario, "choosing the left (right) neighboring category" may be looked at as a previously established convention, much like the "keep to the right side of the road except possibly for passing" convention that evolved for regulating road traffic in various places

being a basis of other conventions, e.g., skating on sidewalks or passing people on an escalator. The adoption of "choosing the left (right) neighboring category" convention in our simulations produced signaling systems that were qualitative different in their dynamics from those that adopted the null convention of choosing randomly.

Non-stationary conventions like those observed for the populations of smoothing and reinforcement learners behave *locally* like conventions based on almost equilibria—that is, behave like conventions based on almost equilibria for appropriate intervals about times t, where t is some time after the convention has been established—but *globally* behave differently from almost equilibria in that the conventional meaning of signals changes with time.

Not all meanings in a non-stationary signaling system need to drift with time. Consider the case of a population signaling game with the color terms *yellow*, *orange*, *red*, *purple*, and *green*. Suppose, as in the simulated population learning game involving smoothing learners and colors from a hue circle, a non-stationary convergent solution is reached where the meanings for the color terms are also organized in the same circular order as the colors they name, say, counterclockwisely as: *yellow*, *orange*, *red*, *purple*, and *green*. Then the proposition that *orange* is immediately between *yellow* and *red* conventionally holds globally, even though the meanings of the individual color terms change with time. This proposition can be viewed as an example of what Lewis (1969) calls "a consequence of the convention." However, because the above convention changes with time, it is better called a *global* consequence of the convention, in order to distinguish it from consequences that only hold *locally*, that is, from consequences that only hold for specific periods of time.

Conventions are at the heart of concepts like social contracts, norms, and conformative behavior. To our knowledge formulations of conventions that allow for drifting meaning have not appeared in the literature, although such "drifting conventions" appear to be important in the modeling of some forms of social and institutional change.

6. Summary

Our intent for this article is to examine categorization methods for color chips that continuously perceptually vary along a circle or a line. In such situations categorization is achieved by playing a repeated evolutionary game involving color naming. We investigated cases where the chips are arranged according to jnd gradients and the game's evolutionary dynamics employ simple learning algorithms and simple rules for determining successes in the games played. Two types of learners are considered:

(1) An individual learner with learning updates that depend only on (i) the presented chips, (ii) the similarity range (i.e., k-similarity), and (iii) his current categorization strategy. This type of learner acquires a categorization system based entirely on individual experience.
(2) A population learner with learning updates that depend, like an individual learner, on (i) the presented chips, (ii) the similarity range, and (iii) his current categorization strategy, but with updating additionally depending on (iv) the current categorization of another randomly chosen learner's categorization of the presented chips and also possibly on that learner's fitness. A population learner acquires a categorization system that is based on both his individual experience and the experiences of other members of the population.

Our results show the emergence of optimal categorization systems across a variety of games played with homogeneous and inhomogeneous stimuli, for individual agents and across individual agents in a population. In these categorization systems the success rate is maximized, and each category has a unique name. Our simulations showed that the following generally holds:

- In both individual and population learning, learners categorize poorly if memoryless learning algorithms are employed.
- In both individual and population learning, learners produce a near optimal categorization system in which category meanings can drift if smoothing learning algorithms are employed.
- In both individual and population learning, learners produce a near optimal categorization system in which category meanings do not drift if reinforcement learning algorithms are employed.
- In population learning, learners converge to essentially the same near optimal categorization.
- In population learning, incorporating agent "fitness" does not have an effect on population categorization.
- When the similarity range is constant for all chips, all categories are of approximately the same size; when stimuli with varying similarity ranges are considered, categories of different sizes evolve.
- When an inhomogeneous color distribution ("color diet") is considered, the most frequently sampled regions in the stimuli space develop color categories first.
- Categories may "drop out," i.e., the categorization that emerges through evolution may develop variable sized and fewer numbers of categories depending on the similarity ranges of chips. Similarly, color terms that occur very infrequently in the population at first may become adopted by the entire population, thus increasing the total number of color terms, in response to similarity requirements.

The main conclusion of this article is that a few simple hypotheses about color discrimination combined with learning through a simple language game can reproduce several general findings in the empirical literature

concerning color naming within a population. To reproduce finer features such as the prevalence of many naming systems with blue–green categories, or evolutionary splitting of a category, or schemes that classify known color-naming systems, and so on, additional hypotheses about perceptual color organization and additional algorithms that take into account more complicated, pragmatic, social interactions are needed. Forthcoming research investigates whether this can be accomplished using only a simple form of reinforcement learning.

Acknowledgments

The authors gratefully acknowledge Brian Skyrms, Ragnar Steingrimssom, Rory Smead, Kevin Zollman and Kate Longo for insights and suggestions during the development of this project. Helpful comments on the manuscript were made by Galina Paramei, David Bimler, Delwin Lindsey, Angela Brown, and anonymous reviewers. We thank Bruce MacEvoy for permission to use the illustration in Fig. 1. Portions of this research were presented at the Society of Mathematical Psychology annual meeting (Jameson, Komarova & Narens, 2006). Part of this research was supported by NSF#0724228. N.K. gratefully acknowledges support of a Sloan Fellowship.

References

Beggs, A. W. (2005). On the convergence of reinforcement learning. *Journal of Economic Theory*, *122*(1), 1–36.

Belpaeme, T., & Bleys, J. (2005). Explaining universal color categories through a constrained acquisition process. *Adaptive Behavior*, *13*, 293–310.

Börgers, T., & Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, *77*, 1–14.

Boynton, R. M. (1997). Insights gained from naming the OSA colors. In C. L. Hardin (Ed.) *Color categories in thought and language*. (pp. 135–150). Cambridge, UK: Cambridge University Press.

Cross Cultural Research (2005a). Special issue on color categorization. *Cross Cultural Research*, *39*(1), 5–106.

Cross-Cultural Research (2005b). Special issue on color categorization. *Cross Cultural Research*, *39*(2), 111–227.

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Sciences*, *4*, 258–267.

Garner, W. R. (1974). *The processing of information and structure*. Hillsdale, New Jersey: Lawrence Erlbaum.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *64*, 377–388.

Grice, P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Griffin, L. D. (2006). The basic colour categories are optimal for classification. *Journal of the Royal Society: Interface*, *3*, 71–85.

Hardin, C. L. (1988). *Color for philosophers*: *Unweaving the rainbow*. Indianapolis, IN: Hackett.

Hardin, C. L. (2005). Explaining basic color categories. *Cross-Cultural Research*, *39*, 72–87.

Hardin, C. L., & Maffi, L. (1997). *Color categories in thought and language*. UK: Cambridge University Press.

Hauser, M. D. (1996). *The evolution of communication*. Cambridge, MA: Harvard University Press.

Jameson, K. A. (2005a). Sharing perceptually grounded categories in uniform and nonuniform populations. Commentary on Steels, L. & Belpaeme, T. (Target Article). Coordinating perceptually grounded categories through language. A case study for colour. *Behavioral and Brain Sciences*, *28*, 501–502.

Jameson, K. A. (2005b). Semantic and perceptual representations of color. In J. S. Monahan, S. M. Sheffert, & J. T. Townsend (Eds.), *Fechner day 2005: Proceedings of the 21st annual meeting of the international society for psychophysics* (pp. 125–132). Mt. Pleasant: Central Michigan University Printing Services.

Jameson, K. A. (2005c). The role of culture in color naming research. *Cross-Cultural Research: The Journal of Comparative Social Science*, *39*, 88–106.

Jameson, K. A. (2005d). Culture and cognition: What is universal about the representation of color experience? *The Journal of Cognition & Culture*, *5*, 293–347.

Jameson, K. A. (2005e). Why GRUE? An interpoint-distance model analysis of composite color categories. *Cross-Cultural Research*, *39*, 159–194.

Jameson, K. A., & D'Andrade, R. G. (1997). It's not really red, green, yellow, blue: An inquiry into cognitive color space. In C. L. Hardin, & L. Maffi (Eds.), *Color categories in thought and language* (pp. 295–319). UK: Cambridge University Press.

Jameson, K. A., Komarova, N. L. & Narens, L. (2006). Simulating color category evolution. Presentation at the annual meeting of the society for mathematical psychology. July 31, 2006. Vancouver, B.C.

Journal of Cognition & Culture (2005). Special Issue on Cognition and Color Categorization. *Journal of Cognition & Culture*, *5*(3–4), 265–495.

Kay, P. (2005). Color categories are not arbitrary. *Cross-Cultural Research*, *39*, 39–55.

Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, *100*, 9085–9089.

Kay, P., Berlin, B. & Merrifield, W. (1991). Biocultural implications of systems of color naming. *Journal of Linguistic Anthropology*, *I*(I), 2–25.

Komarova, N. L. (2004). Replicator-mutator equation, universality property and population dynamics of learning. *Journal of Theoretical Biology*, *230*(2), 227–239.

Komarova, N. L., & Niyogi, P. (2004). Optimizing the mutual intelligibility of linguistic agents in a shared world. *Artificial Intelligence*, *154*, 1–42.

Komarova, N. L., Niyogi, P., & Nowak, M. A. (2001). Evolutionary dynamics of grammar acquisition. *Journal of Theoretical Biology*, *209*, 43–59.

Komarova, N. L., & Rivin, I. (2003). Harmonic mean, random polynomials and stochastic matrices. *Advances in Applied Mathematics*, *31*, 501–526.

Kuehni, R. G. (2004). Variability in unique hue selection: A surprising phenomenon. *Color Research and Application*, *29*, 158–162.

Kuehni, R. G. (2005). Focal color variability and unique hue stimulus variability. *The Journal of Cognition and Culture*, *5*, 409–426.

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

Malkoc, G., Kay, P., & Webster, M. A. (2005). Variations in normal color vision. IV. Binary hues and hue scaling. *Journal of the Optical Society of America A*, *22*, 2154–2168.

Matsuno, T., Kawai, N., & Matsuzawa, T. (2006). Color classification in chimpanzees (Pan troglodytes). In T. Matsuzawa, M. Tomonaga, & M. Tanaka (Eds.), *Cognitive development in chimpanzees* (pp. 317–329). Tokyo: Springer.

Matsuzawa, T. (1985). Colour naming and classification in a chimpanzee (Pan troglodytes). *Journal of Human Evolution*, *14*, 283–291.

Newhall, S., Nickerson, D., & Judd, D. (1943). Final report of the OSA subcommittee on spacing of the Munsell colors. *Journal of the Optical Society of America*, *33*, 349–385.

Niyogi, P. (1998). *The informational complexity of learning*. Boston: Kluwer.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

Niyogi, P., & Berwick, R. C. (1996). Learning from triggers. *Linguistic Inquiry*, *27*, 605–622.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of universal grammar. *Science*, *291*, 114–118.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, *417*, 611–617.

Regan, B. C., Julliot, C., Simmen, B., Viénot, F., Charles-Dominique, P., & Mollon, J. D. (1998). Frugivory and colour vision in Alouatta seniculus, a trichromatic platyrrhine monkey. *Vision Research*, *38*, 3321–3327.

Regier, T., Kay, P., & Cook, R. S. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences*, *102*, 8386–8391.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*, 1436–1441.

Saunders, B. A. C., & van Brakel, J. (1997). Are there nontrivial constraints on colour categorization? *Behavioral and Brain Sciences*, *20*, 167–228.

Skyrms, B. (1996). *Evolution of the social contract*. Cambridge: Cambridge University Press.

Smith, J. M., & Harper, D. (2003). *Animal signals*. Oxford: Oxford University Press.

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, *24*, 99–142.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories: A case study for colour. *Behavioral and Brain Sciences*, *28*, 469–529.

Steels, L., & Vogt, P. (1997). Grounding adaptive language games in robotic agents. In P. Husbands, & I. Harvey (Eds.), *Proceedings of the fourth European conference on artificial life*. Cambridge, MA: MIT Press.

Stoner, K. E., Riba-Hernández, P., & Lucas, P. W. (2005). Comparative use of color vision for frugivory by sympatric species of platyrrhines. *American Journal of Primatology*, *67*, 399–409.

Zuidema, W., & Westermann, G. (2003). Evolution of an optimal lexicon under constraint from embodiment. *Artificial Life*, *9*, 387–402.