**The origin of color categories**

Daniel J. Garside1#, Audrey L.Y. Chang1,2, Hannah M. Selwyn1, Bevil R. Conway1,3,#

1Laboratory of Sensorimotor Research, National Eye Institute

2 present address: Vilcek Institute of Graduate Biomedical Sciences, New York University.

3National Institute of Mental Health

#correspondence to [danny.garside@nih.gov](mailto:danny.garside@nih.gov) or [bevil@nih.gov](mailto:bevil@nih.gov)

**A characteristic feature of the human mind is the use of concepts{Carey, 2009 #18670;Tenenbaum, 2011 #17954}. To what extent does concept formation depend on language and culture? Here we exploit color as a tool of cognitive science{Lara, 2014 #9993;Wilken, 2004 #18702;Hardman, 2017 #18701;Witzel, 2018 #16170;Schurgin, 2020 #17999;Kim, 2021 #18700;Maule, 2023 #18696;Block, 2023 #18672} and ask if macaque monkeys have a substrate for color concepts, evident as color categories. Color is continuous yet readily categorized, and for humans, color categories express meaning-laden concepts. For example, a ripe strawberry, an embarrassed child, a valentine’s heart, all derive meaning from redness. Macaques have the same spectral sensitivity as humans{Schnapf JL, 1987 #363} and very similar central visual pathways{Lafer-Sousa, 2016 #17582}, but for our purposes they usefully lack language. We adapted a nonverbal match-to-sample paradigm used in humans to uncover color categories from errors in matches{Bae, 2015 #9761}; the colors are sampled from a presumed perceptually uniform space of colors{Brainard, 2010 #4684}. With this method, macaques seemed to show two consensus color categories, not four as in humans; these corresponded to “warm” and “cool”. We next determined whether the results were better explained by simulated data modeling the impact of true cognitive categories or unwitting distortions in the color space. The results strongly favor the second explanation, providing evidence of a nonlinear mapping between a true underlying perceptually uniform representation of colors and the color space. From the empirical results, we reverse-engineered a uniform perceptual space, to our knowledge the first color similarity space unconfounded by language. The results show that human color categories are likely not innate and suggest that widely assumed “uniform” color spaces are inexorably shaped by universally relevant human categories of warm and cool{Gibson, 2017 #9992}.**

Color categories are identified by color terms, of which the Basic Color Terms are considered prominent {Berlin, 1969 #1165}. One hypothesis is that among the BCTs, red, green, blue, and yellow are universal {Heider, 1972 #2762;Regier, 2005 #8218} and endowed by hard-wired neural mechanisms present at birth {Bornstein, 1976 #18689;Lindsey, 2006 #8246}. This idea, put forth 150 years ago {Hering, 1875 #18572}, predicts observed cross-cultural color naming patterns {Jameson, 2009 #7488;Baronchelli, 2010 #8500;Lindsey, 2015 #8219;Abbott, 2016 #8474} and is consistent with some neurophysiological results {Clifford, 2009 #18682;Holmes, 2009 #18683;Brouwer, 2013 #6550;Bird, 2014 #8271;Yang, 2016 #18693;Forder, 2017 #10003}. Behavioral work in infants provides perhaps the strongest evidence for a biological origin of color categories {Franklin, 2004 #18692;Ozturk, 2013 #6777}; this work suggests that the innate categories may be defined by retinal cone-opponent mechanisms rather than basic colors {Skelton, 2017 #14994;Maule, 2019 #18691}. Another hypothesis is that color categories emerge in development, instructed by language and culture {Roberson, 2005 #8197;Regier, 2009 #8250;Cibelli, 2016 #18503}, and possibly involving an interplay of innate and developmental factors {Kay, 2006 #18704;Franklin, 2008 #4562;Regier, 2009 #8250}. This hypothesis is promoted by variability in color naming patterns across languages and individuals {Davidoff, 1999 #6901;Roberson, 2000 #6899;Paramei, 2018 #16202;Webster, 2002 #8200}. Current consensus is that some aspect of adult color category behavior is acquired through experience; but the extent to which the origin of color categories is innate remains unresolved {Davidoff, 2009 #18684;Skelton, 2023 #18695}.

One approach to the origin of color categories that sidesteps difficulties working with human infants is studies of trichromatic non-human primates. The few studies on this topic have come to different conclusions: one found color categories in macaques consistent with categories in human adults {Sandell, 1979 #210}; one tested for a blue-green category boundary and found it in humans but not in baboons {Fagot, 2006 #18502}; and one found different categories in the two macaques tested, apparently dependent on the animals’ experiences {Panichello, 2019 #18694}.

**Measuring color categories in macaque monkeys**

Addressing the question of color categories in monkeys confronts several challenges. First, how to measure color categories without teaching the animals the categories {Essock, 1977 #18698;Matsuno, 2004 #18697}. Second, how to specify the color stimuli {Siuda-Krzywicka, 2019 #17168}; for example, specifying the colors as wavelengths {Sandell, 1979 #210} is not appropriate {Davidoff, 2010 #18699}. Third, how to obtain precise data across the full circle of hues. A match-to-sample paradigm using colors defined in a perceptually uniform color space (**Figure 1a**) provides a potential solution {Bae, 2015 #9761;Panichello, 2019 #18694}. But to test for consistent color categories across animals, we need not only to ensure precision in the matched colors but also to avoid the possibility of reinforcing biases acquired while the animals perform the task. So, rather than having animals match a cued color to a spot on a continuous ring of colors and occasionally rewarding them for inaccurate matches as in the established paradigm, we adapted it as an alternative-forced-choice task in which a direct match to the cue was available in every trial and the monkeys were only rewarded for making the direct match (**Figure 1b**). One consequence of the adapted paradigm is that it requires considerable data to satisfactorily sample category performance across the space of colors. Four animals performed the task, completing at total of 299,690 trials over 232 sessions (**SI Figure 1**).

If a monkey has a color category, the category will be captured by the negative slope in a plot of the choice bias (**Figure 1c**). The approach is data driven so it will recover whatever categories exist; nonetheless, before collecting the data we considered three possibilities. First, that the monkeys would show no color categories, as predicted by the work in baboons {Davidoff, 2010 #18699} (**Figure 1d, top**); second, that macaque color categories would correspond to the four main basic color categories, consistent with data in human adults {Bae, 2015 #9761} (**Figure 1d, middle**); and third, that macaque color categories would align with the cone-opponent mechanisms predicted by data in human infants{Skelton, 2017 #14994} (**Figure 1d, bottom**). The animals performed well on the task, showing an average lapse rate on the easiest trials of X% (**Figure 2a;** plots of individual animals in **SI Figure 2**) and providing clear evidence of choice biases (**SI Figure 3**). But the recovered color categories analyzed with a mixture model do not support any of the predictions (**Figure 2b**). Instead, the animals appeared to show two consensus color categories. These categories are not obviously aligned with either of the cone-opponent mechanisms (arrowheads, **Figure 2c**) but could instead be described as “warm” and “cool” (teal-colored and salmon-colored radiating wedges in **Figure 2c;** data for individual animals is shown in **SI Figure 4**).

**Two possible explanations for choice biases in macaque monkeys**

The colors we used were defined by the International Commission on Illumination (CIE) to be approximately perceptually uniform. But it has long been recognized that there may be non-uniformities in the space {Brainard, 2010 #4684}; some have argued that perceptual uniformity may be task dependent or simply unattainable {Judd, 1970 #17915}. One might even suppose that if language influences color perception, as stipulated by the Sapir-Whorf hypothesis, then all color spaces generated by human observers could be shaped by language. Could the macaque consensus color categories be attributed not to a true “cognitive” category (**Figure 3a**) but to unrecognized distortions in the presumed uniform space of colors (**Figure 3b**)? The central difference in the two explanations can be understood by considering the relationship between two neighboring colors. For the cognitive-bias account, there is an asymmetry between the colors if there is a category center nearby. The color further from the category center will be more likely mistaken for the color closer to the category center than the other way around. Whereas for the non-uniform color space, there is no asymmetry in mismatches between neighboring pairs of colors.

To illustrate that the behavioral data could be caused by either a cognitive-bias account or a stimulus-space non-uniformity, we generated two sets of simulated color matching data. One simulation was derived from a uniform space with two color category centers and the other was derived from a distorted color space with two foci of distortion. Both simulations give rise to the same pattern of results when analyzed with a mixture model, and the pattern qualitatively matches the behavioral data (**Figure 3c,d**; compare these simulations with Figure 2c). So, to tease apart the possible underlying causes of the behavioral results, we developed a generative model based on the Target Confusability Competition (TCC) model {Schurgin, 2020 #17999}. The key modification of our model is that it does not assume the same underlying similarity function for each color. Instead, the shape of the similarity function can be adjusted to discover the structure underlying the choice biases (by contrast, the TCC model uses an exponential function with additional perceptual noise). We refer to our model as TCC for recovering choice biases (TCC-c) and illustrate the output with a similarity matrix that we describe as “free”, since the similarity function of each color is free to be optimally fit independently of the functions for all other colors, including those next to it. The simulated data for the cognitive-bias account was generated by imposing an asymmetric relationship among pairs of neighboring colors, and the TCC-c model recovers that structure as an asymmetry in the similarity matrix about the diagonal (**Figure 3e**). The simulated data for the stimulus space non-uniformity account, meanwhile, was generated by imposing symmetric relationships between neighboring pairs of colors but varying the relationship distance between pairs around the color wheel. The TCC-c model recovers this structure as a symmetric bulge about the diagonal (**Figure 3f**).

To determine which model better explains the macaque behavioral data, we quantitatively compared the free similarity matrix of the macaque behavioral data (**Figure 4a, b**) with TCC-c models that were constrained to fit using either the cognitive-bias similarity functions, the stimulus space non-uniformity similarity functions, or a combination. The behavioral data are much better explained by a model that reflects nonuniformities in the color space than by a model that reflects cognitive biases, as quantified by negative log-likelihood and AIC (**Figure 4c,d**). In fact, the cognitive-bias model does no better than the null hypothesis. These results strongly suggest that macaque monkeys do not have innate color categories. And if the macaque is an accurate model of the human, then the results imply that humans do not have innate color categories either.

But does the TCC-c model have sufficient sensitivity to discover a cognitive bias should one exist? The data for individual animals shows that it does. By mixture-model analysis, one animal showed not only the two consensus choice biases for “warm” and “cool” but also a strong choice bias for pea green (**Figure 5a**). The free similarity matrix for the data from this animal shows a asymmetry about the diagonal corresponding to this color (**Figure 5b**), providing the hallmark of a cognitive color bias that is confirmed by negative log-likelihood and AIC quantification (**Figure 5c, d**).

**A perceptually uniform color space unconfounded by language**

The behavioral data in macaques provide a rare opportunity to reconstruct a perceptually uniform color space unconfounded by language. We computed, empirically, the matrix required to transform the spacing of stimuli such that the macaques would, on average, show no choice bias. When colors evenly sampled from the purportedly uniform CIELu\*v\* space (**Figure 6a**) are plotted within this macaque-derived uniform color space, colors around the teal part of the space collect, and to a lesser extent, so do colors in the salmon-peach part of the space (**Figure 6b**).

The non-uniformities in the CIE u\*v\* space implied by these results raise a question about their origin. Given color categories can be learned, as evident in at least one macaque in our study (**Figure 5**) and two macaques in another study {Panichello, 2019 #18694}, coupled with the likely possibility that this learning reflects colors of environmental or behavioral relevance, we wondered whether the distortions in the presumed uniform color space could be attributed to the warm and cool color categories manifest in apparently all human cultures {Gibson, 2017 #9992} and hypothesized to be caused by the non-arbitrary color statistics of objects and backgrounds {Rosenthal, 2018 #15658}—objects are more likely to be warm colored and backgrounds, cool colored. To test this idea, we ran a simulation in which the macaque uniform color space was used in a color-matching task by an agent with cognitive biases for warm and cool colors defined by the average colors of objects and backgrounds. The resulting similarity space of colors shows a distortion—a non-uniformity—that corresponds to the distortion inferred from the macaque behavioral data. These results provide a plausible mechanism by which the distortions in the color space arise and suggest that all color spaces made by human observers are inexorably impacted by the universally behaviorally relevant categories of warm and cool.

Left to incorporate:

different colorspaces, optimised for different tasks, could exist at different levels of the visual hierarchy.

Saturation bias

4-Alternative Forced Choice: Human participants

**Methods**

*Subjects*

Data were collected in four adult male rhesus macaques (*Macaca mulatta*)(“PO, CA, BU, and MO”) weighing 8–10 kg. All experimental procedures were approved by the Animal Care and Use Committee of the National Eye Institute and complied with the regulations of the National Institutes of Health. Plastic headposts were mounted with sterile surgical procedures, using procedures described in detail elsewhere {Lafer-Sousa, 2013 #6318}. The animals were acclimatized with positive reinforcement to sit in a custom-made chair positioned with the eyes 57 cm in front of a computer monitor and to perform visual tasks as described below.

*Behavioral task*

At the beginning of each testing session we positioned a mouthpiece to deliver fluid reward to the animal. The animals were trained to perform a 4-Alternative Forced Choice (4-AFC), Delayed Match to Sample task. Each trial was initiated when the animal fixated a small symbol at the center of the screen; trials were aborted if the animal did not maintain fixation until the fixation spot disappeared. Eye position was monitored with an infrared camera (ISCAN), and fixation was defined as within a ~1o wide area centered on the fixation spot. The trial sequence was as follows. Fifty ms after initiating the trial by fixating the central spot, a 2-degree-diameter “cue” pseudo-randomly drawn from the set of 64 colors appeared for 750ms at a location within 3-degrees of the fixation spot. The position of the cue was consistent for a given daily recording session and could vary from [where in the visual field]. The cue was followed by a gray screen for a brief “memory” period of 600-900ms, after which four match options appeared at the same eccentricity from the fixation spot, evenly spaced apart, in the visual hemifield opposite to the cue, with the exact positions of the choices varying randomly trial-to-trial. One choice option was always a direct match to the cue, and the other three were randomly sampled without replacement from the remaining 63 stimuli. After a random period from 500 to 1000 ms, the fixation spot disappeared, instructing the monkey to direct its eyes to one of the choices. Reward was given only if the animal selected the choice that was identical to the cue. If the monkey failed to make a choice within 5 seconds, the trial aborted. Animals made choices on average within 250 ms (95% CI). The cue colors of aborted trials were randomly re-introduced into the sequence of future trials so that on average the trials included about the same number of cue colors, with the same difficulty of selection across colors. Task difficulty was defined by the set of foil colors that accompanied the direct match for the match options: trials were considered more difficult for smaller differences between the cue color and the foil closest to the cue color. The experiment was controlled with custom software written in MATLAB and Psychtoolbox {Kleiner, 2007 #4658}.

*Stimuli*

Colors were defined by an equiluminant plane in CIELUV color space, and were of equal saturation within this space. Stimuli were discs presented on a Cambridge Research Systems Display++ screen under neutral adapting conditions (adapting field was xyY). The color of the discs varied only in hue (not luminance contrast or saturation), and were sampled from 64 equally spaced points on a circle in CIELUV space with a white point of XXX (xy), a radius of 37, and a luminance of XXX (L\* = 76.0693). These values were chosen to maximize gamut while maintaining constant saturation and luminance. CIELUV was used, in contrast to previous work which has used CIELAB, because CIELUV has the benefit of an associated chromaticity diagram. Luminance noise was added by XXX to the extent of YYY.

*4-Alternative Forced Choice: Human participants*

Human participants were recruited via Amazon Mechanical Turk to perform an analogous version of the non-human primate 4-AFC task. The purpose was simply to test the extent to which the modified paradigm, providing for matches to one of four colors, recovers similar results to those obtained using a paradigm where matches are made to a ring of all the colors. To request a trial, participants used a mouse to adjust the location of a cursor to click on a fixation cross, after which a cue was shown to one side of the fixation cross and the cursor disappeared. The cue was displayed for 750 ms. After the cue was extinguished, a fixation cross was shown but the cursor remained hidden to de-incentivize mouse movement (1500 ms). Four choices were then shown and the cursor reappeared, and participants made their selection by using the mouse to move the cursor to their choice and clicking the mouse. The pattern of results (**SI Figure 6**) was qualitatively similar to those in published studies using the continuous ring of colors; like the published studies {Bae, 2015 #9761;Panichello, 2019 #18694}, the choice biases were best modeled with a four-category model.

*Data Analysis**: Mixture Modeling*

To assess the bias in responses for each cue, we computed the distribution of responses on trials where the monkey made an incorrect choice. For each completed trial, we calculated the error as the angular difference between the correct option and the chosen option. For each cue, we computed the number of times the monkey selected each incorrect choice, normalized by the number of times each choice color was available as a choice option for all completed trials of the given cue (this was approximately uniformly distributed). We then fit a Gaussian with a variable floor:

(Equation 1) ([[eq:GaussianEquation]](#eq:GaussianEquation))

to the error distribution for each cue, using the MATLAB *fit* function with the equation defined as:

(Equation 2) a\*exp(-(((x-b)^2)/(2\*c^2)))+d.

$$\label{eq:GaussianEquation} \eqnmarkbox[purple]{p1}{a} \cdot \exp \frac{-(x- \eqnmarkbox[violet]{mu}{\mu} )^2}{2 \eqnmarkbox[blue]{sigma}{Gilbert, #6818}^2} + \eqnmarkbox[gray]{d}{d}$$

This fit was weighted by the number of times each choice color was an option for the given cue across all completed trials. Bias was taken as the difference between the cue and the peak of the corresponding Gaussian, for each cue color (, Equation 1). The values, for each stimulus, are plotted as the black lines in [[fig:BiasCurvesCombined]](#fig:BiasCurvesCombined). Where this line falls closer to the center of the figure than the 0 line, there is negative bias (which in this representation is counter-clockwise), and vice versa for values above the 0 line. Confidence intervals were extracted from ??

*Data Analysis: Modified Target Confusability Competition Model (TCC-c)*

One disadvantage of the mixture model for our analysis is that we can only use it to analyze the subset of trials where the animal made an incorrect response. This is a consequence of the 4-AFC paradigm, where across all trials, correct choices will be over-represented; there is no straightforward way to establish chance level when including trials with correct choices. In order to use the full dataset (incorrect and correct trials), we developed a generative model, based on the Target Confusability Competition (TCC) model of reference {Schurgin, 2020 #17999}. The key elements of the TCC model are a similarity function, which determines the similarity between stimulus and stimulus through a non-linear mapping of distance to similarity, and a value of , which can be thought of as describing the amount of noise acting in the system. These two elements can be used to predict the probability that a choice of colour will be picked from the set of , on a trial where the cue is . Our implementation of the model includes some important modifications required for our purpose.

First, our model does not assume that the underlying function is the same for each stimulus. The standard TCC model obtains a single similarity function averaged across stimuli, but it does not have to, as Schurgin et al demonstrate (see Figures 1D and Extended Data Figure 5 of ref.{Schurgin, 2020 #17999}). Since we are most interested in the differences between the functions for the different stimuli, we allow the similarity function to be fit independently for every color. Second, our model makes no assumptions about the underlying function that determines similarity. The standard TCC model uses an exponential function with additional perceptual noise (see Figure 1F of ref. {Schurgin, 2020 #17999}), based on observations gained from collecting data on various simultaneous judgment tasks. Our model does not do this because we expect that, if choice biases are present, they will require a differently shaped similarity function for different stimuli to be discovered. Third, we refer to fits made this way as “free” fits (and the corresponding similarity matrix as “free”), since each element of the matrix is “free” to be fit optimally, independently of all other colors. Fourth, we fit our model on single datasets at a time, whereas the standard TCC model derives the similarity functions and values for from independent datasets.

In fitting a free similarity matrix, noise in the system can either be represented by the value of or by modifying the “contrast” of the similarity matrix (the relationship between the highest values and the lowest values in the matrix), since we apply no constraints on the floor or peak of the function. We therefore assume a value of for the similarity matrix fits.

Since we use an AFC method, as opposed to a pseudo-continuous response space, we are able to take advantage of an alternative computational method for computing the probabilities of a particular choice being made. We use the correction factors of (McGraw and Wong 1992) (their Table 3) to estimate the probability , where are samples from independent normal variables, with means representing the pairwise similarity values between and , and variances determined by . This decreases the runtime of our model by several orders of magnitude compared to the standard TCC method of ref. {Schurgin, 2020 #17999}. (See the function modelPDF in TCC\_Code\_InManuscriptOrder\Model\TCCUncorrelated.m from <https://osf.io/j2h65/> for comparison).

Our modification of the TCC model, which we refer to as TCC-c (TCC modified to recover choice biases), uses parameter estimation techniques to construct a similarity matrix where each cell represents the similarity between stimulus and stimulus . This model, and this visualization method, allows us to assess not only the mean of the bias in the responses, but the shape of the response-bias-curve, which provides insight into the source of the biases.

*Determining Cognitive Bias vs. Stimulus-Space Non-Uniformity using TCC-c*

For these experiments, we used a nominally perceptually uniform color space: CIELUV. This space has been derived psychophysically, with the goal of minimizing differences in perceptual non-uniformity across the space, for color differences of small magnitudes (the apparent color difference between two points in one part of the space should be equal to the apparent color difference between two points in another part of the space when that the cartesian distance between the two points in each case be the same).

However, non-uniformities within the space are thought to exist {Brainard, 2010 #4684}; and uniformity for small color differences does not assure uniformity for larger color differences {Judd, 1970 #17915}. Moreover, uniformity for the conditions under which the psychometric measurements from which the space was determined does not assure uniformity across all viewing conditions {Siuda-Krzywicka, 2019 #17168}. Non-uniformities in color space could lead to choice biases that manifest in a similar way to true cognitive biases when analyzed with a mixture model (see **Figure 3**). Distinguishing between the alternative sources of the choice biases can be accomplished with the TCC-c model analysis. We define “true cognitive bias” as when an agent is more likely to pick choice as a match to cue than they are to select choice as a match to cue , and that this behavior would not be reciprocal (they would not be more likely to pick choice as a match to cue than they are to select choice as a match to cue ). By this definition, this type of bias would appear as an asymmetric spread away from the negative diagonal in the similarity matrix; symmetry across the diagonal would represent reciprocity.

Variations in the symmetric spread around the negative diagonal in the similarity matrix would represent non-uniformity of the stimulus space. In areas where the behavioral space is oversampled, one would see spread away from the negative diagonal (adjacent colors are more similar than the average). In areas that are undersampled, one would see a pinch into the negative diagonal (adjacent colors are less similar than the average).

To estimate how much of the bias in the similarity matrix of the macaque behavior data could be attributed to non-uniformity of color space, we fit a version of the TCC-c model using a single similarity function, defined by

(which is controlled by two parameters: and , which together control the slope of the function and the extent of the flat-top of the function at zero) and instead allow the stimuli chromaticities to float. These two parameters can be thought of as the *similarity function* (how similar is to ), and the *perceptual function* (at what point do stimuli become indistinguishable from one-another). Unfortunately, in this parameterization of the function, the parameters are highly correlated, which makes recovery of these values via model fitting difficult. An alternative parameterization where the parameters were maximally uncorrelated would be preferable. This is akin to asking: what set of relationships between the stimuli in stimulus space can best explain the data we observe?

$$\label{eq:SimilarityFunction} \eqnmarkbox[purple]{explambda}{\exp(x\cdot\lambda)} \eqnmarkbox[cyan]{convolution}{\circledast} \eqnmarkbox[blue]{Gilbert, #6818}{\mathcal{N}(0,\sigma^2)}$$

Could choice biases be attributed to un-intended variation in saturation across the colors? This seems unlikely because we observed little or no bias towards higher saturation colors in a control experiment (SI FIGURE?). Still, it is also seen that the number of errors of the inverse type (upper right quadrant) is roughly equal in number.

*Reverse-engineering a uniform color space from the macaque color-matching data*

Following the fitting of the model described above, the set of modified stimuli chromaticities can be extracted ([[fig:MACBEHspace]](#fig:MACBEHspace). These can be thought of as the chromaticity values for the stimuli that we used, no longer represented in CIELUV, but now in a behaviorally derived color space instead. It is then possible to define a uniformly distributed set of colors in this new space, reparameterize them by their relationship to the original colors in the new space (e.g. new chromaticity is 60% of the angular distance on the path between stimulus and stimulus ), and then determine where this new set of colors would be in CIELUV ([[fig:UniformStimsInCIELUV]](#fig:UniformStimsInCIELUV)).

Data availability statement

Data is available at: zenodo(?)...

Code availability statement

Code as used in this paper: doi... Latest version of code: github...

Acknowledgments

Author contributions

References