Effects of Category Addition to the Perceptual Magnet Effect

John Andrew Chwe

Introduction

The perceptual magnet effect described in the seminal Bayesian model paper (Feldman & Griffiths, 2007) is founded on the categorization of the sounds /i/ and /e/. The conventions of the English language are such that stability exists surrounding the perceptual distributions of these sounds in the physical spectrum of Hz; although the distributions representing /i/ and /e/ sounds will vary between perceiver in both category mean Hz and variance, we can expect these distributions to be generally similar across perceivers. English as a language does not incentivize the creation of a 3rd phonetic category in-between the /i/ category mean and the /e/ category mean. Such structure is not innate and is dependent on the stimulus space one inhabits. For example, classic work on the discrimination of speech sounds demonstrated that American participants show categorical discrimination between pairs of sounds on a *rah-lah* continuum, whereas Japanese participants performed around chance-level (Miyawaki et al., 1975).

The impact of one's environment on category structure has also been found in the social domain.

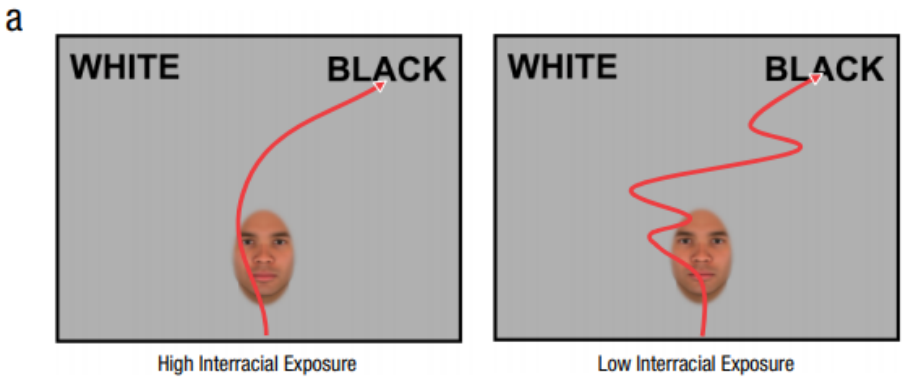

Figure 1. Mouse trajectories for participants with high interracial exposure show smoother decision pathways compared to low interracial exposure participants.

In a MouseTracker task, individuals with a low level of interracial exposure demonstrated more unstable mouse trajectories when categorizing a racially ambiguous face compared to those with high level of interracial exposure, suggesting that the social categories of white and black race are more differentiated for low-exposure participants compared to the high-exposure participants (Freeman, Pauker, & Sanchez, 2016).

However, categories in the social and non-social domain likely do not share the same level of variance across perceivers. For example, in the aforementioned study, levels of interracial exposure were approximated using Census data and participant's self-reported zip code. In such a situation, the category representations of race vary with a large amount of granularity even within a given state, whereas category representations of a drill or a car are likely consistent across countries. Other recent work in the emotion and trait perception domain have illustrated how idiosyncratic beliefs about the overlap between emotion and trait categories impact the perception of such in faces (Brooks & Freeman, 2018; Stolier et al., 2018). For example, individuals who believe two emotion categories overlap to a high degree produce more similar visual representations in a reverse correlation task (Brooks & Freeman, 2018). Collectively, emerging research provides evidence of substantial variance in category structure in social domains.

Given that category structure differs between individuals, I am interested in how pre-existing category structure impacts the effects of a new category. I will explore the effects of a category addition on the perceptual magnet effect under two pre-existing category structures (narrow/wide). As a concrete example, take for consideration the case where an individual grows up in a rural, highly conservative environment. In this hypothetical environment, the maintenance of a strict gender binary is a societal priority. As such, there is little variance in what is considered male or female, and this rigidity is reflected in gender presentations in the

community (i.e. stimuli space). As a result, individuals in this environment possess a representation of gender with two distributions (male/female) with small variances. Also of interest, but not explored here, is the fact that individuals in this hypothetical, generic environment find no distinction between gender and sex.

In contrast, take for consideration an individual who grew up in a less extreme environment, and what it means to be male or female has a broader, less specific definition. In this context, the individual still possesses a representation of gender with two distributions, but the variance of the distributions is wider. I will explore how the addition of a third category impacts the amount of perceptual magnet effect found in situations with a narrow or broad pre-existing category structure.

Returning to the aforementioned example, this addition of a category could be likened to learning about the category of "gender non-conforming," or having increased exposure to non-prototypic gender expressions through some form of experience, motivating the creation of a third gender category to accommodate individuals who are neither clearly male or female. I will use this framing in the following sections, although it is of more practical use rather than theoretically relevant. I am interested in this broadly, in both social and non-social contexts.

The impact on the perceptual magnet effect/categorical perception is important as the phenomenon can be seen as a dimension-reduction technique, where excessive categorical perception reduces the nuance of the raw world that can provide adaptive individualization to decision-making and precurse stereotyping/prejudice. Specifically, I am interested in how the addition of a third category impacts the amount of perceptual magnet effect depending on existing category structure.

Methods

The exploration was conducted using the existing framework from Feldman & Griffiths (2007). Specifically, I created 40 "faces" that exist in 0.5 unit increments from -10 to 30 on a hypothetical unidimensional gender dimension. Two gender categories were created with means of 0 and 10. For the sake of this exploration, I will call these categories "Male" and "Female", although the framing of this is not of great importance. These categories were assigned equal prior probabilities of 0.5. For the narrow condition, these categories were assigned standard deviations of 0.5. For the wide condition, standard deviations of 2 were assigned. Note that all categories must have equal standard deviations for the formulas given in Feldman & Griffiths (2007) to hold.

In both the wide and narrow 2 category scenarios, I then used the following formula to calculate the posterior probability that a category produced a perceived face. This is referred to category identification by Feldman & Griffiths (2007).

$$p(c|S) = \frac{p(S|c)p(c)}{\sum_c p(S|c)p(c)}$$

The likelihood p(S|c) is defined as the following: $p(S|c) = \int p(S|T)p(T|c)\, dT$. Using this calculation, I then found the expected value of T given S using the following:

$$E[T|S] = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_S^2} S + \frac{\sigma_S^2}{\sigma_c^2 + \sigma_S^2} \Sigma_c p(c|S)\mu_c$$

I then added a third category to both the wide and narrow 2 category scenarios. This third category was assigned a mean of 5. Note that the added category's standard deviation must equal that of the existing categories for the provided formulas to hold. The added category was given a

prior of 0.2, with both pre-existing category priors being adjusted to 0.4 as a result. Finally, I evaluated the total amount of magnet effect occurring in all models as the absolute value of the difference between the stimuli and E[T|S].

Results

Below are the category identification graphs for all four scenarios:

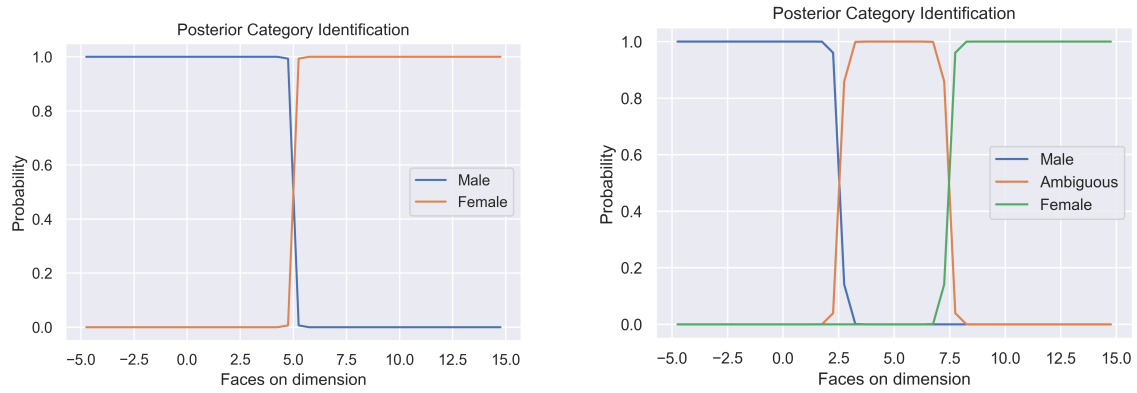

Figure 2. Category identification graphs for the narrow 2→3 scenario (0.5 SD).

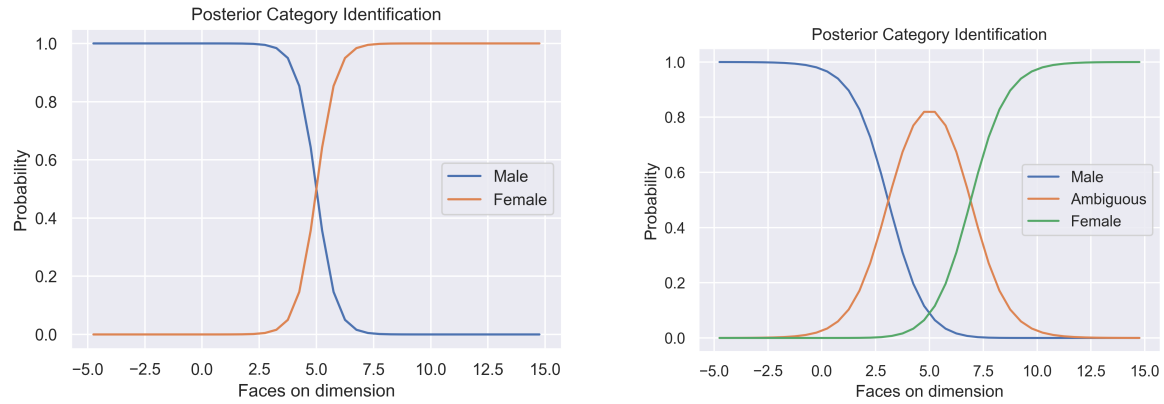

Figure 3. Category identification graphs for the wide 2→3 scenario (2 SD).

Unsurprisingly, the posterior probability of category membership is more categorical in the narrow condition than in the wide condition, and the wide, 2 category identification graph is consistent with that found by Feldman & Griffiths (2007). Below are visualizations of the perceptual magnet effect in both the narrow and wide 2 category scenario.

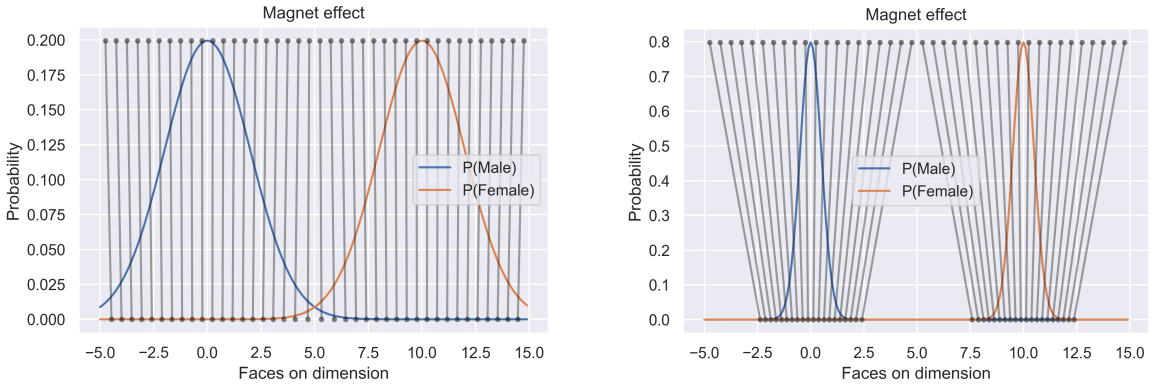

Figure 4. Visualizations of the perceptual magnet effect in both the wide and narrow 2 category scenarios, respectively.

I was able to successfully replicate the same findings of Feldman and Griffiths, where "manipulating category variance yields extreme categorical perception in categories with low variance and perception that is less categorical in categories with high variance (p. 261)." The observed perceptual magnet effects are quite slight for the wide scenario, while being quite apparent in the narrow scenario. Following, I then created the same graphs for the narrow and wide 3 category scenarios.

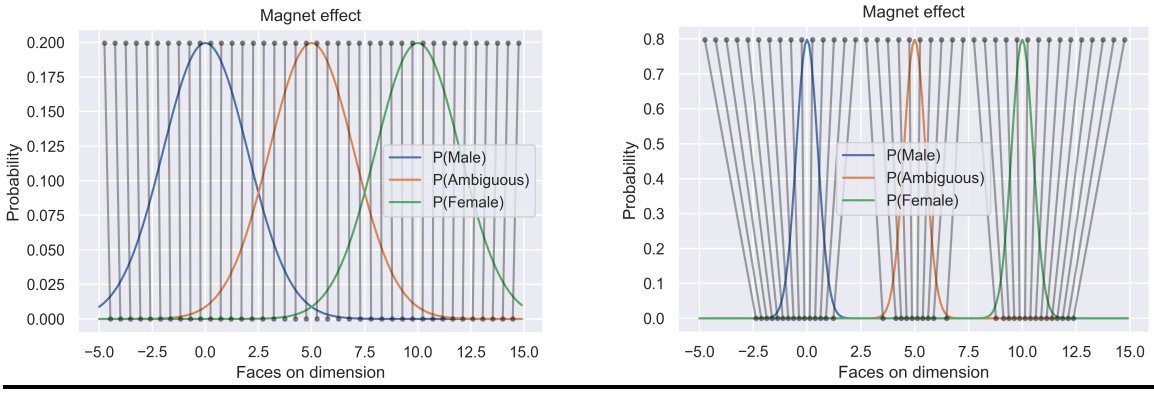

Figure 5. Visualizations of the perceptual magnet effect in both the wide and narrow 3 category scenarios.

With the addition of a third category, the perceptual magnet effect in the wide condition truly becomes scarce. In the narrow condition, points previously equidistant from both category means

are now drawn to the new category mean, decreasing the amount of perceptual magnet effect seen.

To quantify the changes in perceptual magnet effect from the 2 to 3 category scenarios, I calculated the absolute value of the difference between E[T|S] and each corresponding "face", shown in the graph below.

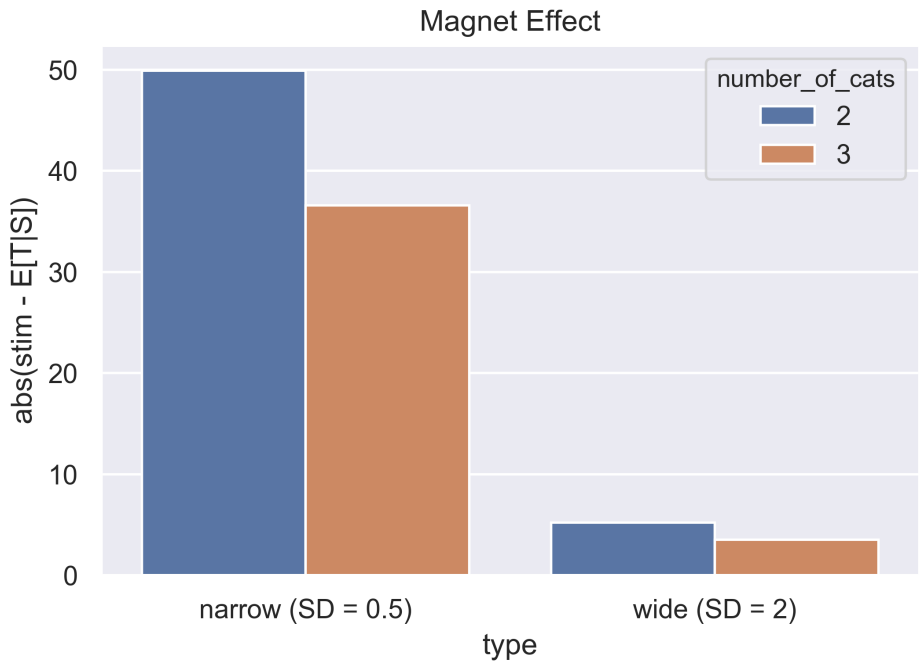

Figure 6. The sum of the absolute difference between each "face" and its corresponding E[T|S] per condition.

Notably, the absolute change in perceptual magnet effect when adding a third category is larger for the narrow condition than the wide condition. However, when analyzed as a percent change, the addition of a third category led to a 26.72% decrease in magnet effect in the narrow condition and a 32.81% decrease in the wide condition. It seems fairly apparent from the magnet visualizations that the decrease in magnet effect seen can be attributed to faces that were experiencing extreme magnet effect under the 2 category scenario that are better accounted for, so to speak, under the 3 category scenario. In order to assess this more formally, below are

distributions of the amount of perceptual magnet effect in all scenarios. Note that all histograms have been normalized such that the probability mass of each sums to 1.

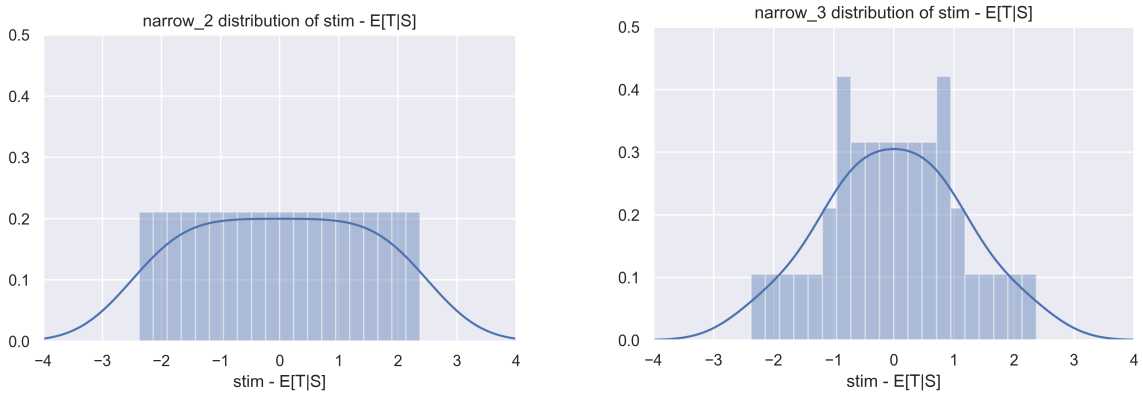

Figure 7. Distributions of perceptual magnet effect in the narrow scenario.

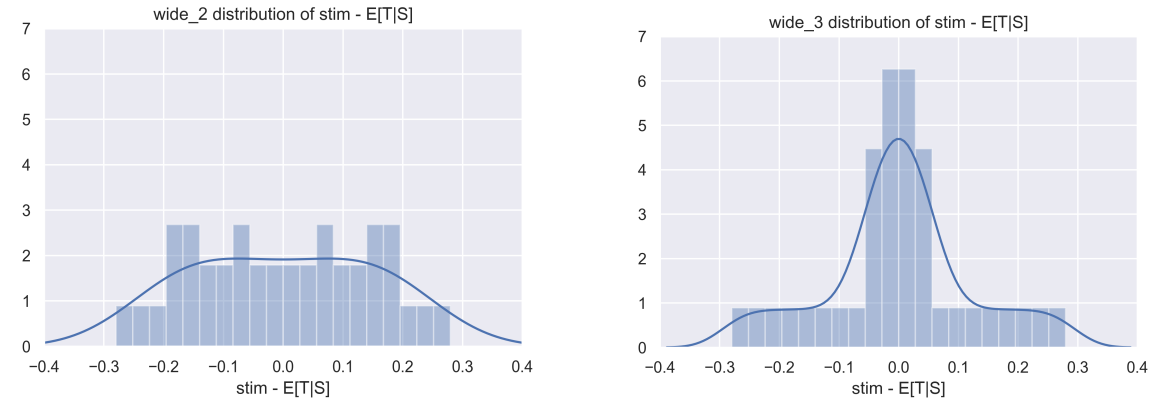

Figure 8. Distributions of perceptual magnet effect in the wide scenario.

As expected, the decrease in total perceptual magnet effect can be attributed to extreme cases of the magnet effect being decreased. This can be seen in the increase in normality of the magnet effect distributions when adding a third category.

Discussion

In the above work, I explored how the effect on categorical perception of adding a third category is dependent on pre-existing category structure. The addition of a third category decreased the amount of perceptual magnet effect in both the narrow and wide scenarios. The absolute decrease in magnet effect was largest for the narrow condition than the wide condition, although the percent decrease was larger for the wide condition. In both the wide and narrow conditions, this decrease can be reasonably linked to a reduction in the number of "faces" experiencing extreme categorical perception.

I am interested in the addition of a category and its impact on perceptual magnet depending on pre-existing structure as an exploration on how increasing the nuance of a representation decreases the inaccuracy of perception and the associations of a representation. Explicitly, "increasing the nuance of a representation" in the given hypothetical case refers to adding a category to gender (male/female → male/ambiguous/female) and "decreases the inaccuracy of perception and the associations of a representation" refers to the decrease in magnet effect. The amount of perceptual magnet effect here can be seen as a proxy of category structure, as they are clearly intertwined, and classic work in social psychology has linked personality trait needs for "simple structure" to stereotype use (ex. Neuberg & Newsom, 1993). In the above simulation, I'd liken the decrease in perceptual magnet effect by introduction of a third category as an increase in how reflective the category representations are of the "ground truth" world.

Given that world exists in an exceptionally high dimensional space, a space far too complex to operate in, it is clearly necessary to simplify our representations. This is categorization canon. But where does the optimum balance between nuance/complexity and parsimony exist in our representations?

The value of highly nuanced, high dimensional representations was very evident this past election cycle, where Cambridge Analytica combined traditional ad targeting with psychometrics and big data. Wielding hundreds of data points from personality type to magazine subscriptions, Cambridge Analytica meticulously sought to individualize advertisements and maximize their effect. From a VICE article released in January of 2017:

> On the day of the third presidential debate between Trump and Clinton, Trump's team tested 175,000 different ad variations for his arguments, in order to find the right versions above all via Facebook. The messages differed for the most part only in microscopic details, in order to target the recipients in the optimal psychological way: different headings, colors, captions, with a photo or video. This fine-tuning reaches all the way down to the smallest groups, Nix explained in an interview with us. “We can address villages or apartment blocks in a targeted way. Even individuals.” (Hannes & Grassegger, 2017).

This is an effect of accuracy through nuance. The effect is a product of both having enough data to produce nuanced representations and the ability to do so. For the given goal of influencing voters, being able to categorize voters with a high level of nuance led to greater accuracy in ad impact, as each category has its own unique associations.

Clearly, the level of nuance created by Cambridge Analytica in categorizing voters is impossible in humans. Yet, given that the world exists in a dimensionality far beyond human capacity, the broader a generalization or representation, the more likely it is to be wrong (ex. stereotypes). Only more nuanced representations have the potential to be more accurately reflective of the “ground truth.” For example, if I know an individual is a man, then I have different predictions about their personality than if I did not have that data point. We can argue that this is a 1-dimensional representation. If I know an individual is a man who is a professor,

then I have further changed predictions about their personality. If I know an individual is a male professor of math, the expectation changes once again. And if I know a male professor of math who I've lived with for 5 years, further nuancing my representation of them as an individual, at what point does my representation of them yield a personality expectation that fully converges with his "ground truth" personality? There is certainly a tradeoff between parsimony and nuance, and at a certain point, there is a decrease in the value of more nuance, but this optimality is not at a 1-D representation.

Perception holds an important role in this process, as perception provides the amount of data, so to speak, that we have available to create representations with. As shown previously, both the level of perceptual capacity and the space of things to be observed impact category structure. This influence has also been shown to be cross-modal. Russian as a language provides distinction between light and dark blue, and this difference has been associated with increased perceptual discrimination between blues compared to English participants (Winawer, 2007).

Moving forward, collecting human data would clearly emerge as the next step. Such data would provide various parameters to the model (ex. category means, variances, etc.), as well as a validation for the model output as seen in Feldman & Griffiths (2007).

In adjusting the model, I would be interested in how to adjust the provided formulas to accommodate categories with different variances, as well as expanding the model into higher dimensionalities. In the current model, "gender" is a unidimensional construct, which is likely simplistic. Models with 2 dimensions are highly popular in psychology, from stereotype content to person perception. Expanding this Bayesian model to 2 dimensions would have immediate application.

Existing models of person construal posit a recurrent relationship between high-level stereotypes and categories and low-level processing of visual cues. The model argues that these

top-down and bottom-up factors interact dynamically to create a final person construal (Freeman & Ambady, 2011). A similar question can be asked here. How does perceived nuance/dimensionality impact category structure, and how does such structure in turn impact our perception?

In sum, the addition of a third category decreased the perceptual magnet effect observed in both the narrow and wide conditions. Future work would use human data to model this change in magnet effect and evaluate how additional categories impact decision-making and stereotyping. This Bayesian framework provides an interesting perspective through which to explore the above questions.

References

Brooks, J. A., & Freeman, J. B. (2018). Conceptual knowledge predicts the representational structure of facial emotion perception. *Nature Human Behaviour*, *2*(8), 581-591.

Feldman, N. H., & Griffiths, T. L. (2007). A rational account of the perceptual magnet effect. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 29, No. 29).

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological review*, *118*(2), 247.

Freeman, J. B., Pauker, K., & Sanchez, D. T. (2016). A perceptual pathway to bias: Interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychological Science*, *27*(4), 502-517.

Grassegger, H. & Krogerus M. The Data That Turned the World Upside Down. VICE News. https://www.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win.

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, *18*(5), 331-340.

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of personality and social psychology*, *65*(1), 113.

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*, *115*(37), 9210-9215.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780-7785.