Explanatory Biases in Social Categorization

Samuel G. B. Johnson[1], Haylie Shestle Kim[2], & Frank C. Keil[1]
(samuel.johnson@yale.edu, hayliekim@berkeley.edu, frank.keil@yale.edu)
[1]Department of Psychology, Yale University, New Haven, CT 06520 USA
[2]Department of Linguistics, University of California, Berkeley, CA 94720 USA

Abstract

Stereotypes are important simplifying assumptions we use for navigating the social world, associating traits with social categories. These beliefs can be used to infer an individual's likely social category from observed traits (a *diagnostic* inference) or to make inferences about an individual's unknown traits based on their putative social category (a *predictive* inference). We argue that these inferences rely on the same *explanatory logic* as other sorts of diagnostic and predictive reasoning tasks, such as causal explanation. Supporting this conclusion, we demonstrate that stereotype use involves four of the same biases known to be used in causal explanation: A bias against categories making unverified predictions (Exp. 1), a bias toward simple categories (Exp. 2), an asymmetry between confirmed and disconfirmed predictions of potential categories (Exp. 3), and a tendency to treat uncertain categorizations as certainly true or false (Exp. 4).

Keywords: Social categorization; inductive reasoning; stereotyping; explanation; causal reasoning.

Introduction

Stereotypes help us to navigate the social world. Like other categories, social categories allow us to use a subset of an individual's properties to predict that individual's other properties (Murphy, 2002). If a cat has stripes, it may be a tiger and therefore aggressive. If a person has a law degree, she may be a lawyer—and therefore litigious.

However, stereotypes often underdetermine what inferences to draw, even in the absence of individuating information. Angie, for instance, is both a woman and a Texan. What should we predict about her pool hustling skill? Our stereotype of Texans suggests she should do well, whereas our stereotype of women suggests she should do poorly. That is, social categories are *cross-classified* (e.g., Ross & Murphy, 1999): Individuals can belong to multiple categories simultaneously, so the outcome of stereotype-based inference can be ambiguous.

A related complication is that it is often unclear what social categories individuals belong to. This causes ambiguities in using stereotypes both in diagnostic (trait-to-category) and predictive (category-to-trait) reasoning. For instance, Sarah is socially awkward and fascinated by sci-fi novels. While these traits fit well with our stereotype of engineers, it is also possible that she is (for example) a lawyer. That is, the diagnostic inference about Sarah's occupation is uncertain. This uncertainty also propagates to predictive inferences about her other traits—she is more likely to be good at math if she is an engineer and more likely to sue you if she is a lawyer.

A great deal is known about *when* stereotypes are applied to individuals (e.g., Hilton & von Hippel, 1996). However, less is known about *how* people use stereotypes to make inferences about individuals when multiple social categorizations are plausible—a problem exacerbated by the prolific cross-classification of human beings.

We propose that stereotype-based inferences follow *explanatory logic*—a set of heuristics used to perform diagnostic and predictive inferences across various types of psychological processes (e.g., Johnson, Rajeev-Kumar, & Keil, 2015; Johnson, Jin, & Keil, 2014). Put differently, social categories are not merely statistically associated with stereotypical traits; instead, people seek out the category that best *explains* an individual's traits, and do so using a specific set of domain-general heuristics. This proposal is a specific version of the broad theoretical orientation that emphasizes attributional processes (e.g., Hamilton & Sherman, 1996; Pettigrew, 1979), explanatory theories (McGarty, Yzerbyt, & Spears, 2002; Wittenbrink, Hilton, & Gist, 1998), and intuitive schemas (Fiske, 1993; Hilton & von Hippel, 1996) in stereotyping. (See also Tversky & Kahneman, 1983 for related evidence of heuristic processing in stereotype use.)

Here, we look for signature biases of explanatory logic in stereotype-based social categorization. Experiments 1–3 look at diagnostic reasoning biases (the role of *inferred evidence*, a *simplicity heuristic*, and an *asymmetry in positive vs. negative evidence*). Experiment 4 looks at a predictive reasoning bias (*digitizing* uncertain beliefs).

Experiment 1: Inferred Evidence

Suppose you have been recruited by your friend to help determine the identity of a cocktail she has been served, which is spicy, orange-flavored, and hazel-colored. Is it mixed with triple sec (orange-flavored and clear) or Grand Marnier (orange-flavored and hazel-colored)? The hazel color of the orange-flavored liqueur would be the deciding factor. Unfortunately, the spicy flavor means that the drink contains spiced rum—which is also hazel-colored. In the absence of further evidence, there is just no way to tell which is the right liqueur.

However, people do not settle for ignorance in such cases—they try to *infer* what color the liqueur was, and use erroneous cues to do so, such as the base rate of liqueurs that are clear versus hazel-colored. They may conclude that most liqueurs are clear, therefore this particular liqueur is likely clear, therefore triple sec is the probable culprit (Johnson, Rajeev-Kumar, & Keil, 2015). This logic is incorrect because it relies on a reference class that is too broad: People ought to consider only the

drinks made with either triple sec or Grand Marnier (of which 50% would be made with hazel-colored liqueurs); instead, they consider *all* drinks (of which most are made with clear liqueurs). This bias typically leads people to infer explanations that make fewer unverified predictions (Khemlani, Sussman, & Oppenheimer, 2011).

Such reasoning extends far beyond the bar. It appears in causal explanation (Khemlani et al., 2011), classification (Sussman, Khemlani, & Oppenheimer, 2014), and causal strength judgments (Johnson, Johnston, Toig, & Keil, 2014), and emerges early in childhood (Johnston, Johnson, Koven, & Keil, 2015). Thus, inferred evidence appears to be used across diverse explanatory tasks.

Do people also use this strategy when using stereotypes to reason about individuals? For instance, if you meet someone highly educated, is she likelier to be an engineer or a lawyer? If you focus on her math skills (unclear from the conversation), you may conclude she is likely a lawyer. But if you focus on her argumentativeness (also unclear), you may conclude she is likely an engineer.

To test this bias, participants were told about either the personality traits (Exp. 1A) or physical traits (Exp. 1B) of an individual from an unfamiliar culture. They were told that the individual had one trait (e.g., hard-working) but that it was unknown whether the individual had another trait (humorous). Participants were told about stereotypes for two groups, one of which was associated with both the known and unknown trait (a religion whose members are known to be hard-working and humorous) and the other associated with only the known trait (an occupation category whose members are known to be hard-working). If people infer that unknown traits are likely to be absent (based on low base rates), then they should categorize her (non-normatively) into the one-trait group.

Method

We recruited 200 participants from Amazon Mechanical Turk for Experiment 1A (N = 100) and Experiment 1B (N = 100); 18 were excluded from analysis due to poor performance on check questions (see below).

Participants were instructed to:

> Suppose you are visiting a foreign country called Gazda. The people of Gazda belong to many different kinds of groups, including different religions, occupations, and ethnicities.
>
> During your visit, you hear about some citizens of Gazda, but you aren't sure which groups they belong to. However, you do have some information from your friend, a native Gazdan. Your task will be to try to figure out which groups these citizens belong to.

Next, participants completed three items, concerning inferences based on either personality (Exp. 1A) or physical (Exp. 1B) stereotypes. For example:

> You've heard that Taylor is hard-working, but no one has told you whether or not Taylor is humorous.
>
> There are two groups of people who are known to be hard-working:
>
> About 1 in 10 people belong to the religion of Ghalism, and they have a reputation for being hard-working and humorous.
>
> About 1 in 10 people have the occupation of Chener, and they have a reputation for being hard-working.

Participants then categorized the individual ("Which of the following groups do you think Taylor is more likely to belong to?") on a 0 ("Ghalism religion") to 10 ("Chener occupation") scale. The category order was randomized.

The personality traits (Exp. 1A) were either positive (hard-working, humorous), negative (dishonest, arrogant), or neutral (traditional, emotional). The physical traits (Exp. 1B) were not valenced (brown bracelets, blue shoes, tall hats, black clothes, left ear piercings, white bottomed shirts). Items were presented in a random order.

At the end of each study, a set of check questions were asked. Participants incorrectly answering 30% or more of these questions were excluded from analysis.

Results and Discussion

Measures were recoded so that negative scores (-5 to 0) correspond to the one-trait category and positive scores (0 to 5) correspond to the two-trait category.

Participants in Exp. 1A showed a significant bias toward the one-trait categories [M = -0.91, SD = 1.28; $t(93) = 6.92$, $p < .001$, $d = 0.71$, $BF_{10} > 1000$], which was consistent across personality traits that were positive [M = -0.97, SD = 1.72], neutral [M = -0.81, SD = 1.71], and negative [M = -0.95, SD = 1.65]. Participants in Exp. 1B showed a similar bias for physical traits [M = -0.84, SD = 1.63; $t(88) = 4.84$, $p < .001$, $d = 0.52$, $BF_{10} > 1000$]. The effect size did not differ across experiments [$t(180) = 0.46$, $p = .64$, $d = 0.05$, $BF_{01} = 7.8$].

These results demonstrate a non-normative bias in social categorization, stemming from a more general explanatory heuristic. Social stereotypes are rich with associated attributes (Andersen & Klatzky, 1987) so that some potentially diagnostic traits are sure to be unknown at a given time for an individual. Thus, future work might test whether inferences about natural social categories depend (erroneously) on which unknown traits are salient.

Experiment 2: Simplicity

Your friend is on her second drink, which is nutty and strong. It could be amaretto (both nutty and strong), or it could be Frangelico (nutty but not strong) mixed with rum (strong but not nutty). You probably find the amaretto explanation more plausible, because it is simpler: Before knowing anything about the taste, it is more likely. You would not be alone: For deterministic causal systems, people are biased toward simple explanations, because people use simplicity as a heuristic for estimating prior probabilities (Lombrozo, 2007). That is, people assume that simple explanations are more often true than complex explanations, even overriding the actual base rates.

However, while complex explanations typically have lower prior probabilities, they are often better fits to the

data. Keep in mind that rum, Frangelico, and amaretto do not *always* taste nutty or strong. Thus, we must estimate not only the prior probability of each explanation, but also how likely the actual taste is, given each explanation. A combination of multiple ingredients may be more likely to lead to a strong and nutty taste than one ingredient alone.

Hence, the *complexity* of an explanation can be used to estimate its fit to the data, even as its *simplicity* is used to estimate its prior probability (Johnson, Jin, & Keil, 2014). One piece of evidence for this *opponent heuristic* model is that the simplicity bias is stronger for deterministic than for stochastic causal systems. The goodness-of-fit is perfect for deterministic causal systems, so complexity does not lead to an explanatory benefit; but the goodness-of-fit is imperfect for stochastic causal systems, affording a benefit to the better-fitting complex explanation.

Some stereotypes are relatively homogeneous (e.g., race and skin color), whereas others are far more heterogeneous (e.g., race and driving habits), in parallel to the distinction between deterministic and stochastic causal systems (e.g., Park & Hastie, 1987). Would we also find a difference in stereotype-based categorization between homogeneous and heterogeneous stereotypes?

Method

We recruited 100 participants from Amazon Mechanical Turk; 24 were excluded from analysis.

Participants read about four sets of traits. For two, the categories were *heterogeneous* relative to the traits:

> You've heard that Taylor is greedy and impatient.
> There are three groups of people who are known to be greedy and/or impatient:
> Most (50% of the people) who believe in the religion of Ghalism have a reputation for being greedy and impatient.
> Most (70% of the people) who have the ethnicity of Folian have a reputation for being greedy.
> Most (70% of the people) who have the occupation of Chener have a reputation for being impatient.

Given this information, the conditional probability of the traits is equal given the simple category (Ghalism) and complex category (Folian and Chener). For the other two items, the categories were *homogeneous* relative to the traits (e.g., "All (100% of the people) who…").

Participants then categorized the individual ("Which of the following groups do you think Taylor is more likely to belong to?") on a 0 ("Ghalism religion") to 10 ("Folian ethnicity and Chener occupation") scale. The category order and content was randomized. Items were presented in a random order, and counterbalanced with condition.

Results and Discussion

Measures were recoded so that negative scores (-5 to 0) reflect simple categorizations and positive scores (0 to 5) reflect complex categorizations.

When the categories were homogeneous, participants were biased to categorize the individuals into simple categories [M = -0.96, SD = 2.59; $t(75)$ = -3.23, $p < .001$, d = -0.37, BF_{10} = 11.3], consistent with previous work using deterministic causal systems (Lombrozo, 2007). But when the traits were heterogeneous, participants had no significant bias [M = 0.41, SD = 2.31; $t(75)$ = 1.55, p = .13, d = 0.18, BF_{01} = 3.4], leading to a difference between conditions [$t(75)$ = 3.66, $p < .001$, d = 0.53, BF_{10} = 39.9].

Although this experiment used personality traits, one might expect that there are generalized expectations about the homogeneity of personality versus physical traits. Indeed, in follow-up experiments we have shown that people prefer simpler categorizations based on physical traits (which are more homogeneous) compared to personality traits (which are more heterogeneous).

These findings have potential implications for intergroup bias, in light of work documenting greater perceived homogeneity in one's outgroup relative to one's ingroup (Park, Ryan, & Judd, 1992). For instance, one may prefer to explain the behavior of an outgroup member in terms of only their outgroup category, whereas people may be more willing to explain ingroup members' behavior using sets of overlapping categories.

Experiment 3: Positive vs. Negative Evidence

It's time for your friend's third drink, and you are assessing whether it is a martini (which is dry) or a G&T (sparkling and dry). If she reports that it is both sparkling and dry, it is probably a G&T, because that explanation can account for both pieces of evidence, whereas a martini could account for only one piece of evidence—the G&T explanation makes more *confirmed* predictions, or has more *positive* evidence in its favor. Conversely, suppose the drink is dry but *not* sparkling. Then, it is probably a martini, because the G&T explanation makes the *disconfirmed* prediction that it is sparkling—the G&T explanation has more *negative* evidence against it.

In general, negative evidence is weighed more heavily than positive evidence (Johnson, Merchant, & Keil, 2015a) in a way that appears to be non-normative. That is, in the case where the drink is dry and sparkling, a G&T will be a somewhat better explanation than a martini (because G&T has more *positive* evidence); but in the case where the drink is dry but *not* sparkling, a martini will be a *much* better explanation than a G&T (because G&T has more *negative* evidence).

Exp. 3 tests whether this bias persists when judging the probability of social categories based on stereotypes, by measuring the relative weight placed on positive and negative evidence. We also sought to replicate Exp. 1 by measuring the weight placed on unknown evidence.

Method

We recruited 196 participants from Amazon Mechanical Turk for Experiment 3A (N = 99) and Experiment 3B (N = 97); 10 were excluded from analysis.

Participants completed four items, concerning religious, ethnicity, occupation, and social class categories. The

stereotyped traits were either personality traits (Exp. 3A) or physical traits (Exp. 3B). Items were assigned to the *Y*, *YY*, *YN*, and *YD* conditions, using a Latin square.

For the *Y* item, participants were given one piece of positive evidence for a categorization:

You've heard that Jamie is traditional.

People who have the occupation of Chener have a reputation for being traditional.

For the *YY* item, an additional piece of *positive* evidence was included, relative to the *Y* item:

You've heard that Jamie is traditional and humorous.

People who have the occupation of Chener have a reputation for being traditional and humorous.

For the *YN* item, an additional piece of *negative* evidence was included, relative to the *Y* item:

You've heard that Jamie is traditional but that Jamie is not humorous.

People who have the occupation of Chener have a reputation for being traditional and humorous.

For the *YD* item, an additional piece of *unknown* evidence was included, relative to the *Y* item:

You've heard that Jamie is traditional but no one has ever told you whether or not Jamie is humorous.

People who have the occupation of Chener have a reputation for being traditional and humorous.

After each item, participants were asked to rate the probability that the individual belonged to the category ("How likely do you think it is Jamie belongs to the occupation of Chener?") on a 0–10 scale.

Results and Discussion

Positive evidence scores were calculated by subtracting ratings of Y from ratings of YY, *negative evidence* scores were calculated by subtracting ratings of YN from ratings of Y, and *unknown evidence* scores were calculated by subtracting ratings of YD from ratings of Y.

For personality traits in Exp. 3A, the effect of negative evidence [$M = 2.82$, $SD = 2.96$; $t(82) = 8.69$, $p < .001$, $d = 1.53$, $BF_{10} > 1000$] was larger than the effect of positive evidence [$M = 0.63$, $SD = 1.60$; $t(82) = 3.57$, $p = .001$, $d = 0.44$, $BF_{10} = 30.3$], leading to a significant difference [$t(82) = 5.30$, $p < .001$, $d = 1.40$, $BF_{10} > 1000$]. Similarly, for physical traits in Exp. 3B, the effect of negative evidence [$M = 3.00$, $SD = 2.35$; $t(85) = 11.87$, $p < .001$, $d = 1.61$, $BF_{10} > 1000$] was larger than the effect of positive evidence [$M = 1.07$, $SD = 2.01$; $t(85) = 4.91$, $p < .001$, $d = 0.61$, $BF_{10} > 1000$], leading to a significant difference [$t(85) = 4.89$, $p < .001$, $d = 0.97$, $BF_{10} > 1000$].

We anticipated significant effects of latent evidence as well, given the bias in Exp. 1. Confirming this prediction, there was a significant detrimental effect of adding unknown evidence both for personality traits [$M = 1.36$, $SD = 1.61$; $t(82) = 7.70$, $p < .001$, $d = 0.74$, $BF_{10} > 1000$] and for physical traits [$M = 1.16$, $SD = 1.98$; $t(85) = 5.45$, $p < .001$, $d = 0.62$, $BF_{10} > 1000$].

These results are consistent both with Exp. 1 and with work on causal explanation and category-based induction (e.g., Johnson, Merchant, & Keil, 2015a). This reaffirms the idea that social categorization relies on the same heuristics as other diagnostic reasoning processes. That said, negative evidence often prompts social perceivers to *subtype* an individual, or create a new subordinate category to accommodate disconfirmations of a stereotype (e.g., Brewer, Dull, & Lui, 1981). Future work might address the interaction of these two inference strategies.

Experiment 4: Belief Digitization

Your friend is *pretty* sure her fourth drink is a Long Island Iced Tea—a sure harbinger of a hangover. But she's not *completely* sure—it could instead be a rum and coke. Suppose the two of you want to engage in the morbid exercise of calculating the probability of a hangover. Let's say she is 70% sure the drink is a Long Island Iced Tea (in which case there is an 80% chance of a hangover) and a 30% chance it is a rum and coke (with a 20% chance of a hangover). If we do the math, there is a 62% chance of a hangover (70% * 80% + 30% * 20%).

However, even when fully sober, people do not reason in this normative way, using *graded* probabilities that respect the fact that it is uncertain which drink she has. Instead, people tend to *digitize* (Johnson, Merchant, & Keil, 2015b; Murphy & Ross, 1994), treating explanations as though they are certainly true or certainly false (i.e., a 100% chance of a Long Island Iced Tea). In that case, the probability of a hangover seems to be much higher (80%).

Exp. 4 tested whether people would similarly use social categories in a 'digital' manner to make predictions about a target individual: If one believes that a person probably, but not definitely, belongs to some social category, does one then behave as though that categorization is *certain*?

Method

We recruited 198 participants from Amazon Mechanical Turk for Experiment 4A ($N = 100$) and Experiment 4B ($N = 98$); 38 were excluded due to poor performance on check questions and 51 due to ratings of P(*A*) and P(*B*,*C*) that did not sum to 100%. Analyses including the latter participants lead to the same conclusions.

Participants completed three items, each including information about the traits associated with three categories (similar to Exp. 2):

People who believe in the religion of Ghalism have a reputation for being business-minded and liberal.

People who have the ethnicity of Folian have a reputation for being business-minded.

People who have the occupation of Keader have a reputation for being liberal.

Then, participants were told that two potential categorizations make different predictions about another trait. For one item, in the *high/low* condition, the simple (hence, likelier) categorization has a high probability of that trait, while the complex (less likely) categorization has a low probability of that trait:

When people believe in the religion of Ghalism, they

are usually formal.

When people have the ethnicity of Folian and the occupation of Chener, they are occasionally formal.

In the *low/low* condition, both categorizations have a low probability of the trait:

When people believe in the religion of Ghalism, they are occasionally formal.

When people have the ethnicity of Folian and the occupation of Chener, they are occasionally formal.

In the *low/high* condition, the simple categorization has a low probability of the trait, while the complex categorization has a high probability:

When people believe in the religion of Ghalism, they are occasionally formal.

When people have the ethnicity of Folian and the occupation of Chener, they are usually formal.

Participants were told that an individual had both of the relevant properties for making a categorization ("You've heard that Taylor is business-minded and liberal").

Following this information, participants completed a *diagnosis* question and a *prediction* question, appearing on separate pages. For the diagnosis question, participants rated the probability of the simple categorization ("Taylor believes in the religion of Ghalism") and complex categorization ("Taylor has the ethnicity of Folian and the occupation of Chener"), and asked to ensure their probabilities added up to 100%. For the prediction question, participants estimated the probability of the additional trait ("What do you think is the probability that Taylor is formal?") on a 0 to 100 scale.

Results and Discussion

For the diagnosis questions, participants in both experiments favored the simple over the complex categorizations. When the categorizations were based on personality traits in Exp. 4A, participants judged the simple category [M = 61.5, SD = 14.8] more likely than the complex category [M = 38.5, SD = 14.9]. Likewise, when the categorizations were based on physical traits in Exp. 4B, participants again judged the simple category [M = 64.0, SD = 18.0] more likely than the complex category [M = 36.0, SD = 18.0]. These simplicity preferences are consistent with Exp. 2 and with previous work on simplicity preferences (Lombrozo, 2007). Importantly, however, the probability of the complex category was non-negligible—38.5% in Exp. 4A and 36.0% in Exp. 4B. These would be quite large probabilities to ignore.

Nonetheless, responses to the prediction questions revealed that participants used only the conditional probability of traits given the simple (high-probability) category, ignoring the possibility that the complex (low-probability) category was correct. Participants in Exp. 4A rated the probability of the additional trait (e.g., formality) higher in the high/low than in the low/low condition [M = 65.4, SD = 20.6 vs. M = 57.1, SD = 24.2; $t(48) = 2.08$, p =.043, d = 0.41, BF_{01} = 1.2]. That is, people used the feature likelihoods given the high-probability category when making predictions about that feature, since manipulating that likelihood (high/low vs. low/low) influenced predictions. However, participants did not use the likelihoods given the low-probability category: The low/high and low/low conditions did not differ [M = 56.6, SD = 23.6 vs. M = 57.1, SD = 24.2; $t(48) = -0.15$, p = .88, d = -0.02, BF_{01} = 8.8]. That is, manipulating the feature likelihood given the low-probability category (low/high vs. low/low) did not influence predictions. Thus, people tacitly 'digitize' high-probability categorizations, treating them as certainly true when making predictions.

This pattern was replicated in Exp. 4B, with physical traits. Again, participants distinguished between the high/low and low/low conditions when making feature inferences [M = 68.2, SD = 20.9 vs. M = 58.8, SD = 27.0; $t(59) = 2.74$, p = .008, d = 0.46, BF_{10} = 3.3], indicating that they used the feature likelihood given the high-probability category. However, they did not distinguish between the low/high and low/low conditions [M = 59.0, SD = 22.5 vs. M = 58.8, SD = 27.0; $t(59) = 0.07$, p = .94, d = 0.01, BF_{01} = 9.8], indicating that they ignored the feature likelihood given the low-probability category.

This finding qualifies any claim that social categories are adaptively useful due to their inductive potency: To the extent that social categories help to make predictions about individuals, they lead us to be overconfident in those predictions. This does not mean that categories are not helpful for navigating the social world, just as the corresponding findings in causal explanation do not show that it is useless to explain anything. Nonetheless, this result helps to show how cognitive mechanisms can contribute to prejudice: Even if a stereotype has a grain of truth, people apply it too zealously, failing to take into account other potential categories that may apply.

General Discussion

We often simplify the social world by using stereotypes, assuming that an individual's traits are consistent with their social category. Yet, it is often unclear what social category an individual belongs to. How, if at all, could we rely on stereotypes in such cases?

Here, we have shown that people are subject to a variety of biases in thinking about uncertain social categorizations. People use erroneous cues to *infer evidence* when diagnostic evidence is missing, leading to a bias against categorizations predicting unknown features (Exp. 1 and 3). People prefer *simpler* categorizations (belonging to one category) over more complex categorizations (belonging to multiple categories), but this tendency is eliminated when the stereotypical features are only heterogeneously linked with their categories (Exp. 2). People weigh *disconfirmed* predictions (negative evidence) more heavily *against* a category than they weigh *confirmed* predictions (positive evidence) in its favor (Exp. 3). And when people categorize an individual as *likely* belonging to a category, they treat that individual as *certainly* belonging to that category, when making

inferences about other features (Exp. 4).

These results help to clarify the mechanisms underlying social judgments and allow us to make two new claims about stereotype use: First, stereotypes act as *explanations* in much the same way that intentions explain behavior and causes explain effects; second, people use a set of *heuristics* to evaluate these explanations, which are shared across superficially distinct psychological processes.

These two claims are linked, and rely on the same underlying logic. Many inferential processes share a common informational structure, wherein hypotheses must be evaluated with respect to some body of data. In principle, these problems could be solved through Bayesian updating, accounting for a hypothesis's prior probability and fit to the data. But in practice, people use a variety of heuristics that (at best) *approximate* Bayesian reasoning, and these heuristics are highly similar across tasks such as classification, causal reasoning, and even some visual tasks. We take these heuristics to be a *signature* of explanatory reasoning and the *mechanism* by which these inferences are made. Thus, finding these heuristics at play in stereotype use is strong evidence that this process is both explanatory and heuristic.

We are currently expanding on this work in two ways. First, we are testing other explanatory biases (e.g., a bias to judge explanations with major behavioral implications to be more likely) in the social domain. Second, we are extending these findings from novel stimuli to more realistic, enriched situations. We look forward to the possibility that this work can help reveal how stereotypes are used in realistic settings, and potentially guide interventions to reduce bias and prejudice.

References

Anderson, S.M., & Klatzky, R.L. (1987). Traits and social stereotypes: Levels of categorization in person perception. *Journal of Personality and Social Psychology*, *53*, 235–46.

Brewer, M.B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology*, *41*, 656–70.

Fiske, S.T. (1993). Social cognition and social perception. *Annual Review of Psychology*, *44*, 155–194.

Hamilton, D.L., & Sherman, S.J. (1996). Perceiving persons and groups. *Psychological Review*, *103*, 336–55.

Hilton, J.L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*, 237–71.

Johnson, S.G.B., Jin, A., & Keil, F.C. (2014). Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Johnston, A.M., Toig, A.E., & Keil, F.C. (2014). Explanatory scope informs causal strength inferences. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Merchant, T., & Keil, F.C. (2015a). Argument scope in inductive reasoning: Evidence for an abductive account of induction. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Merchant, T., & Keil, F.C. (2015b). Predictions from uncertain beliefs. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2014). Inferred evidence in latent scope explanations. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2015). *Sense-making under ignorance*. Under review.

Johnston, A.M., Johnson, S.G.B., Koven, M.L., & Keil, F.C. (2015). Probability versus heuristic accounts of explanation in children: Evidence from a latent scope bias. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Khemlani, S.S., Sussman, A.B., & Oppenheimer, D.M. (2011). *Harry Potter* and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*, 527–35.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–57.

McGarty, C., Yzerbyt, V.Y., & Spears, R. (Eds.) (2002). *Stereotypes as explanations*. Cambridge, UK: Cambridge University Press.

Murphy, G.L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G.L., & Ross, B.H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–93.

Park, B., & Hastie, R. (1987). Perception of variability in category development: Instance- versus abstraction-based stereotypes. *Journal of Personality and Social Psychology*, *53*, 621–35.

Park, B., Ryan, C.S., & Judd, C.M. (1992). The role of meaningful subgroups in explaining differences in perceived variability for in-groups and out-groups. *Journal of Personality and Social Psychology*, *63*, 553–67.

Pettigrew, T.F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, *5*, 461–76.

Ross, B.H., & Murphy, G.L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*, 495–553.

Sussman, A.B., Khemlani, S.S., & Oppenheimer, D.M. (2014). Latent scope bias in categorization. *Journal of Experimental Social Psychology*, *52*, 1–8.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Wittenbrink, B., Hilton, J.L., & Gist, P.L. (1998). In search of similarity: Stereotypes as naïve theories in social categorization. *Social Cognition*, *16*, 31–55.