Conceptual constraints on generating explanations

Zach Horne[1] and Sangeet Khemlani[2]
zachary.horne@asu.edu, sangeet.khemlani@nrl.navy.mil

[1]Arizona State University, Phoenix, AZ 85004 USA
[2] US Naval Research Laboratory, Washington, DC 20375 USA

Abstract

When reasoners explain everyday patterns and observations, they tend to generate explanations based on inherent properties of the observations (Cimpian & Salomon, 2014). Cimpian (2015) and his colleagues hypothesized that inherent properties permit rapid explanation, but the mechanism by which reasoners rapidly build explanations remains unclear. Any given concept may relate to innumerable inherent properties, and no theory explains how reasoners avoid protracted searches through semantic memory. Prasada and colleagues (2013) describe a novel conceptual framework that distinguishes between principled and statistical inherent properties. Here, we argue that the framework can resolve the predicted link between rapid explanation and the inherence bias. Two studies provide evidence that people systematically prefer principled inherent explanations. The finding allows for an integrated, mechanistic account of how reasoners generate explanations in which a preference for inherent explanations emerges from a preference for principled connections.

Keywords: inherence bias, principled connections, explanation, reasoning, dual-processes

Introduction

People have the remarkable ability to rapidly produce explanations of complex observations (e.g., Cimpian & Salomon, 2014; Hussak & Cimpian, 2017). When tasked with explaining everyday patterns in the world, people can quickly come up with plausible explanations of why these patterns come about. For instance, how would you explain why people drink lemonade in summer? The explanation reasoners often generate is that the drink is *tart*, which makes it refreshing. The explanation seems sensible (Hyde & Pangborn, 1978) and simple enough for people to comprehend. Yet, it is neither comprehensive nor accurate. The "tartness" explanation doesn't explain why people drink lemonade and not other tart beverages in the summer. An alternative explanation is that marketing campaigns caused the popularity of lemonade, but people tend not to consider such explanations at the outset. Why do people generate sensible but inaccurate explanations? A recent proposal argues that reasoners build initial explanations from restricted and biased retrievals from semantic memory (Cimpian & Salomon, 2014; Hussak & Cimpian, 2017; Tworek & Cimpian, 2016). Those biases produce memories that reference a concept's inherent properties (e.g., the taste of lemonade) instead of its extrinsic properties (e.g., marketing campaigns). Inherent properties are those that are internal to the concept, whereas extrinsic properties often act on a concept but are not part of its composition. The inherence bias predicts that explanations based on extrinsic properties are often the result of deliberation (Hussak & Cimpian, 2017).

Several lines of research corroborate the inherence bias in explanatory reasoning (Cimpian, 2015). Yet, the account does not explain why reasoners appear to prioritize certain inherent properties (e.g., *tartness*) over others (Strevens, 2014): lemonade is often *cold* and *sweet*, for instance, but reasoners tend not to appeal to those properties in their initial explanations. In what follows, we propose a novel account of why people construct initial explanations from only a subset of the inherent properties in memory. We then describe two studies that corroborate the account.

The representation of conceptual knowledge

To construct explanations, reasoners apply abstract conceptual knowledge to concrete situations (Keil, 2006; Lake, Salakhtutdinov, & Tenenbaum, 2015), and so the construction process depends on retrieving relevant semantic memories. Conventional semantic networks efficiently compute relevance by linking certain concepts, such as *hammer*, to other associated concepts, such as *hard* and *metal* (Jones, Willits, & Dennis, 2015; Rodgers, 2008). But, their efficiency depends on implementing a single type of connection that associates different concepts. Prasada, Khemlani, Leslie, and Glucksberg (2013) argued instead that people represent conceptual knowledge through different connections between concepts and properties: *principled* and *statistical* connections. Table 1 provides a set of linguistic tests that distinguish the two connections between the concept *car* and two of its properties: *four wheels* and *radio*. As the table shows, both principled and statistical connections can license generalizations and probabilistic inferences, but only principled connections license normative expectations, formal explanations, aspectual inferences, and inferences about norms. For instance, if a principled property does not hold for an instance of a concept, then reasoners infer that something is abnormal, e.g., a car that doesn't have four wheels is a defective car, or perhaps not a car at all. Statistical connections do not permit such an inference, e.g., there is nothing abnormal about a radio-less car. As Khemlani, Leslie, and Glucksberg (2012) argue, the semantic differences between the two sorts of connections cannot be modeled by unstructured probabilistic accounts.

Principled connections have implications for social reasoning in general. Prior work suggests that principled properties license generalizations (e.g., "lemonade is tart" is felicitous, but "lemonade is cold" is not; Prasada et al., 2013) as well as default inferences (e.g., "given no information to the contrary, this arbitrary glass of lemonade is probably tart"

seems more plausible than "this arbitrary glass of lemonade is probably cold"; Khemlani et al., 2012). In daily life, people make generalizations along racial, gender, sexual, and religious categories, and they draw default inferences based on those generalizations. In such situations, principled properties may yield inferences that are harmful and fallacious (Leslie, 2013; Rhodes et al., 2012).

Table 1. *Linguistic tests that distinguish principled from statistical connections (see Prasada et al., 2013). The '#' denotes linguistic formulations that are unacceptable.*

	Connection	
Diagnostic expectation	**Principled (car → four wheels)**	**Statistical (car → radio)**
Generic generalizations	*Cars have four wheels*	*Cars have radios*
Probability	*Most cars have four wheels*	*Most cars have radios*
Normativity	Cars are supposed to have four wheels	#Cars are supposed to have radios
Formal explanation	That (pointing to a car) has four wheels because it is a car	#That (pointing to a car) has a radio because it is a car
Aspect	One aspect of being a car is having four wheels	#One aspect of being a car is having a radio
Normality	All normal cars have four wheels	#All normal cars have radios

Nevertheless, while principled connections can lead to problematic social inferences, they may explain how reasoners avoid a protracted search through semantic memory when constructing explanations. We posit that that a given concept should maintain far fewer principled connections than statistical ones, and, as a result, principled connections can explain how people avoid traversing a dense conceptual network of background knowledge (cf. Prasada 2017). This would resolve how the inherence bias can operate rapidly to yield commonsense explanations of novel phenomena. The account predicts that the preference for inherent explanations should interact with the preference for principled connections: people should spontaneously retrieve inherent properties that bear a principled connection to the concept when they construct an initial explanation. As a result, they should prefer principled inherent explanations over statistical inherent explanations. Two studies corroborated the prediction.

Study 1

The study tested the prediction that reasoners should prefer principled inherent explanations to statistical inherent explanations.

Method

Preregistration. The projected sample size and predictions for Study 1were preregistered through Open Science Framework. Experimental scripts, analyses, and data are available at https://osf.io/p7aen/.

Participants. 50 participants participated through Amazon Mechanical Turk for monetary compensation. In a post-experimental questionnaire, two participants self-reported that they were not paying attention, and so their data were excluded from the analyses. However, these exclusions did not materially affect the results of the study.

Design, materials, and procedure. Prior work on principled connections, as well as the inherence bias in explanatory reasoning, used materials that describe simple, everyday observations (e.g., Why do people drink lemonade in summer?; see Cimpian & Salomon, 2014; Prasada, 2017). Although these stimuli are easy to understand, they also concern familiar topics, which makes it more difficult to control the number and kinds of inherent and extrinsic factors people maintain. For a more controlled set of stimuli, we created a set of descriptions of novel "scientific" phenomena. We turned to the scientific domain for two reasons: First, it allowed us to remove the possibility of familiarity accounting for the effects we observed because we could fabricate the scientific scenarios given to participants while maintaining a realistic, believable context. Second, it allowed us to control the number and kinds of factors that could plausibly be thought to explain an observation, because we limited the candidate causes discussed in each vignette. Consequently, participants received vignettes that described scientific investigations, such as a vignette about a study on lithium atoms conducted in a high-altitude location. The vignettes made explicit the source of an inherent property (something internal to lithium atoms) as well as the source of an extrinsic property (something external to lithium atoms, such as the high-altitude location). Participants selected from sentences that were formulated to distinguish principled connections from statistical connections. Principled connections should license expectations that the property is a natural aspect of its concept, whereas statistic connections should license expectations of prevalence (see Table 1).

For each vignette, participants selected the most plausible explanation from a set of four alternatives: half of the explanations concerned inherent properties and the other half concerned extrinsic properties. And half the explanations were formulated as a generalization to describe principled connections ("the nature of X") whereas the other half used a generalized quantifier, "most", to describe statistical connections ("what most X is like"). Hence, participants chose from four explanations: a principled inherent explanation, a statistical inherent explanation, a principled extrinsic explanation, and a statistical extrinsic explanation. The four explanations reflected a 2 × 2 within-participants design. The following is a sample vignette with sample response options:

Chemists working in a lab in Denver, Colorado were investigating the possibility of using lithium atoms to store nitride. The chemists found that when they attempted to store nitride in the atoms, it led to a result that completely defied their expectations. Instead of the atoms storing 95% of the nitride, they stored only 2% of the nitride.

Select the most plausible explanation for the observation.
○ *Something about the nature of lithium nitride.* [principled inherent]
○ *Something about what most lithium nitride is like.* [statistical inherent]
○ *Something about the nature of high altitude locations.* [principled extrinsic]
○ *Something about what most high altitude locations are like.* [statistical extrinsic]

The study randomized the order in which the four options appeared, as well as the order of the six vignettes. The materials for this study can be found here https://osf.io/p7aen/.

The materials provided a conservative test of the prediction: the scientists' observations were described as anomalies (e.g., the result "defied their expectations"), which was designed to make it difficult for participants to prefer a principled connection, because anomalous situations could make it less likely that the outcome was the result of a stable, principled property.

Data analysis. Data were analyzed by performing Bayesian estimation using the probabilistic programming language Stan (Gelman et al., 2014) and predictions were tested by computing Bayes Factors (BFs) on the regression coefficients from the model. Larger BFs indicate that the data are more likely under the alternative hypothesis than under the null hypothesis. Additional information regarding the analyses can be found at https://osf.io/p7aen/.

Results and discussion

Study 1 tested the prediction that participants should show a systematic bias toward principled inherent explanations. Figure 1 shows the proportion that participants selected each of the four explanations, and it indicates that people chose principled inherent explanations four times as often as any other option. To confirm the difference, a Bayesian multinomial random-effects analysis modeled participants' responses. The model treated each participant and each vignette as random effects. Table 2 provides the results of the analysis.

The study revealed that people strongly preferred principled inherent explanations to any other explanation (see Table 2; Bayes factors and parameter estimates were similar under different prior choices). For all six vignettes, participants' selections yielded biases towards choosing principled inherent explanations. And 40 out of 49 participants displayed the bias when their performance on individual vignettes was averaged.

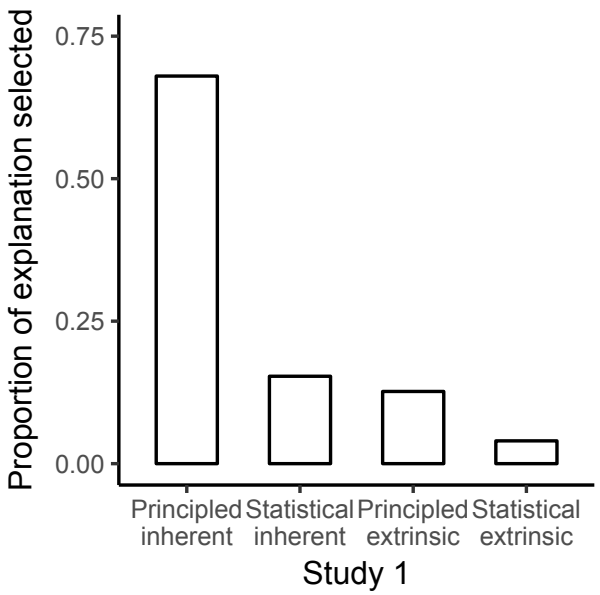

Figure 1. The proportion of explanations selected as most plausible by the participants in Studies 1 and 2.

Study 1 provides evidence for the integrated bias that privileges inherent properties that bear a principled connection to the concept. However, the materials in the study may be confounded in two ways. First, the inherent explanations referenced nouns that were explicitly mentioned in the vignettes (e.g., lithium and nitride) whereas the extrinsic explanations referenced a general property of the environment described in the vignette (e.g., high-altitude) instead of referencing the specific location. Second, the explanations that referenced principled connections used a formulation that involved the phrase "the nature of", which may have been more appealing to participants simply because it matched the domain of scientific investigation and "natural" phenomena. Both of the confounds may have contributed to the effect in Study 1, and so Study 2 eliminated them. It provided participants with both general and specific extrinsic descriptions and it used a formulation for principled connections that made no use of the phrase "the nature of", but nevertheless referred to a generalization (see Table 1).

Table 2. *A multinomial random effects model predicting responses on the basis of random effects of participant and vignette in Study 1. The * indicates that statistical inherent explanations served as the reference in the model.*

		95% CIs		
	Estimate	Lower	Upper	BF_{10}
Statistical inherent*	--	--	--	--
Principled inherent	**1.58**	**.73**	**2.25**	**50.0**
Statistical extrinsic	-1.33	-2.25	-.28	14.3
Principled extrinsic	-0.50	-1.44	.36	.81

Study 2

Study 2 was identical to Study 1 in all respects except for slight variations that eliminated confounds in Study 1.

Method

Participants. 102 Amazon Mechanical Turk workers participated in the study for monetary compensation. An additional participant indicated he or she was not paying attention, and the corresponding data were excluded from analyses.

Materials, design, and procedure. Participants received the same set of vignettes as in Study 1 and they selected the explanation they considered most plausible from a set of four alternatives. The materials, design, and procedure were otherwise the same as in Study 1 with three exceptions. First, participants received an alternative description of principled inherent explanations that made no use of the phrase "the nature of". Instead, participants evaluated principled inherent explanations using the following formulation:

○ *Something about lithium nitride.*
[principled inherent]

Statistical inherent explanations matched those used in Study 1, e.g.,

○ *Something about what most lithium nitride is like.*
[statistical inherent]

Second, Study 2 aimed to rule out the possibility that participants preferred explanations that referred to nouns explicitly mentioned in the vignette. Hence, half of the extrinsic explanations in the study included the location mentioned in the vignette, e.g.,

○ *Something about Denver.*
[extrinsic: location]

and the other half of the extrinsic explanations referred to a salient property of the location, e.g.,

○ *Something about high altitude locations.*
[extrinsic: property of location]

Finally, Study 2 addressed an issue that was peripheral to the question of whether reasoners exhibit a bias toward principled inherent explanations. Specifically, the study tested the assumption that anomalous results should suppress inherent explanations. The study accordingly varied whether the vignettes reference anomalies or not. The materials for this study can be found here https://osf.io/p7aen/.

Results and discussion

The results of Study 2 further corroborated a strong bias towards principled inherent explanations: they were chosen nearly three times as often as statistical inherent explanations (see Figure 2). A multinomial random effects model confirmed the difference; it treated participants' explanation choices as the dependent variable, and it included random intercepts that controlled for variation between participants and vignettes. As in Study 1, the model provided estimates of the difference between the intercept of principled inherent responses and the reference group (i.e., statistical inherent responses; see Table 2).

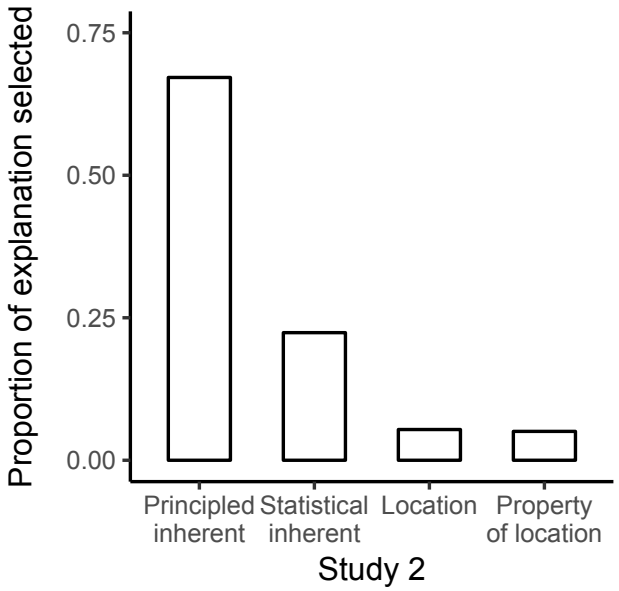

Figure 2. The proportion of explanations selected as most plausible by the participants in Study 2.

Study 2 indicated that the intercept of principled inherent responses was credibly different from the reference group (β = 1.18, 95% CI [.28 to 1.88]), which indicated that people reliably preferred principled inherent explanations over any other explanation type. Here too, we found that for all six vignettes, participants exhibited a bias toward principled inherent explanations. The vast majority of participants (81 of 99) had explanation averages in the predicted direction.

Finally, we found that anomalous findings constrained the bias towards inherent explanations, confirming that describing scientific vignettes as anomalous provided a conservative test of the hypothesis that participants would exhibit a bias towards principled inherent explanations, b = 1.12, 95% CI [.44 to 1.85].

Table 3. *A multinomial random effects model predicting responses on the basis of random effects of participant and vignette in Study 1. The * indicates that statistical inherent explanations served as the reference in the model.*

		95% CIs		
	Estimate	Lower	Upper	BF_{10}
Statistical inherent*	--	--	--	--
Principled inherent	**1.18**	**.28**	**1.88**	**11.1**
Statistical extrinsic	-.85	-2.31	.98	1.62
Principled extrinsic	-.97	-2.78	1.38	1.41

General discussion

When asked to explain a pattern in the world, people quickly generate explanations that oversample the inherent properties of entities in the pattern (e.g., Cimpian & Salomon, 2014). However, extant work has not reconciled how the inherence bias enables the rapid production of an explanation given that numerous inherent properties exist for any given conceptual category. Two studies resolve the question of how the inherence bias enables rapid explanatory inference. They revealed that people preferred inherent explanations that cite properties that bear a principled rather than a statistical connection to the concept under consideration. For example, when asked to explain why a particular group of dogs has multiple mates over a lifetime, people thought it was more plausible that "the nature of" the dog was the cause of the observation than that it was something about "what most [such dogs] are like". Principled connections are rare and privileged (Khemlani, Leslie, & Glucksberg, 2012; Prasada & Dillingham, 2009; Prasada, 2017; Prasada et al., 2013) in that they license generalizations, default inferences, formal explanations, and other sorts of inferences. We posited that concepts have a limited number of principled connections, and that reasoners can efficiently construct explanations by retrieving only those properties that bear a principled connection to a concept. The restriction yields a bias towards inherence, and it curtails a protracted search through semantic memory. Hence, it allows people to rapidly generate explanations.

The present work suggests that an integrated theory of how reasoners generate explanations should couple a preference for inherence with a preference for principled connections, which together enable the rapid construction of explanations. The resulting bias in explanatory reasoning has significant implications that extend beyond controlled vignettes that describe novel scientific observations. Prior research suggests that principled properties permit inferences that statistical properties do not. In many cases, those inferences are sensible: they allow reasoners to accept the generalization that "lemonade is tart, so this glass of lemonade is supposed to be tart" (a default inference), and they reject the same line of reasoning when it concerns statistical properties (e.g., *coldness*; see Khemlani et al., 2012). But if the concept under consideration concerns stereotypes about race or gender, then the inferences permitted by a bias towards principled inherent properties may be problematic (e.g., Leslie, 2013; Rhodes et al., 2012). Reasoners may rapidly explain patterns of individuals on misrepresentative generalizations rather than on external factors. Hence, the present results are consistent with the notion that an explanatory process biased towards principled inherent properties has the downstream consequence of permitting unwarranted generalizations (e.g., Ho et al., 2015) and fallacious inferences (Tworek & Cimpian, 2016).

The present research suggests that people prefer principled inherent explanations over any other explanation type. However, we tested this hypothesis by constructing principled explanations in such a way that may permit a reading that denotes a causal rather than a principled connection. As Prasada et al. (2013) argue, causal connections are implicated in generalizations such as: "sharks attack swimmers." These generalizations appear to concern a causal, dispositional property between, e.g., *sharks* and *attacking swimmers*, such that sharks are disposed to cause the attack to come about. The present studies may concern, not just principled connections, but also causal connections. A major difference between the two is that principled connections license formal explanations, whereas causal connections do not (see Table 1 and Prasada et al., 2013). Future work could therefore test whether people accept formal explanations of the scientific phenomena described above, because formal explanations are diagnostic of principled connections. The present analysis predicts that if participants were asked "Why didn't the lithium atom in this study store only 2% of the nitride?" they should accept the formal explanation: "Because they're lithium atoms." This study would provide a further test of our account of explanatory reasoning.

Another limitation of the present work is that in both studies participants were forced to choose the most plausible explanation. The task required participants to make a definitive judgment, and so it could have exaggerated the extent to which participants preferred principled inherent explanations. A task that does not impose such a constraint could might fail to yield this bias, and thus, future research should examine whether the bias towards principled inherent explanations is robust to different experimental tasks. Still, the present results indicate that participants exhibited a large bias towards principled inherent explanations using new response options and more general wording (Study 2), making it unlikely that the entirety of the effect came from task demands.

A further limitation of the present research is that we tested our account by constructing anomalous events of scientific observations. Although this allowed us to exercise tight control over the stimuli, it narrowed the domain of our materials to the topic of science rather than everyday observations. Furthermore, we tested our hypothesis using only six vignettes. Together, these limitations may limit the generalizability of our findings. We dealt with these concerns, in part, by performing mixed effects modeling and treating vignette as a random effect. It allows us to formally model and generalize to the population of vignettes we could have but did not test. Future studies will examine both more items and domains outside of scientific reasoning to further test the proposed account of explanatory reasoning.

In sum, the robust bias to construct explanations that describe inherent features may coincide with, and emerge from, the underlying representation of the connections between concepts and their properties. One type of connection – the principled connection – appears privileged over others, and the studies we report demonstrate a link between the inherence bias and the bias towards principled connections. The results make progress towards a theory of how explanations are rapidly generated.

Acknowledgments

This work was supported by an NRC Research Associateship Award to ZH and funding from the Office of Naval Research to SK. We thank Andrei Cimpian, Tony Harrison, Laura Hiatt, John Hummel, Joanna Korman, Sandeep Prasada, and Greg Trafton for advice. We also thank Kalyan Gupta, Kevin Zish, and Knexus Research Corporation for their assistance in data collection.

References

Cimpian, A. (2015). The inherence heuristic: Generating everyday explanations. In R. Scott & S. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1–15). Hoboken, NJ: John Wiley and Sons.

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, *37*, 461-480.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian Data Analysis* (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.

Ho, A. K., Roberts, S. O., & Gelman, S. A. (2015). Essentialism and racial bias jointly contribute to the categorization of multiracial individuals. *Psychological Science*, *26*, 1639-1645.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593-1623.

Hussak, L. J., & Cimpian, A. (2018). Memory accessibility shapes explanation: Testing key claims of the inherence heuristic account. *Memory & Cognition*, *46,* 1-21.

Hyde, R. J., & Pangborn, R. M. (1978). Parotid salivation in response to tasting wine. *American Journal of Enology and Viticulture*, *29*, 87-91.

Jones, M. N., Willits, J., Dennis, S., & Jones, M. (2015). Models of semantic memory. *Oxford Handbook of Mathematical and Computational Psychology*, 232-254.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57,* 227-254.

Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2012). Inferences about members of kinds: The generics hypothesis. *Language and Cognitive Processes, 27,* 887-900.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science,* 350, 1332-1338.

Leslie, S. J. (2013). Essence and natural kinds: When science meets preschooler intuition. *Oxford Studies in Epistemology*, *4*, 108-166.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, *2*, 113-162.

Prasada, S. (2017). The scope of formal explanation. *Psychonomic Bulletin & Review, 24,* 1478-1487.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition, 99*, 73-112.

Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: A window onto the formal aspect of common sense conception. *Cognitive Science, 33*, 401-448.

Prasada, S., Khemlani, S., Leslie, S. J., & Glucksberg, S. (2013). Conceptual distinctions amongst generics. *Cognition*, *126*, 405-422.

Rhodes, M., Leslie, S. J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526-13531.

Rogers, T. T. (2008). Computational models of semantic memory. *The Cambridge handbook of Computational Psychology*, 226-266.

Strevens, M. (2014). The causes of characteristic properties: Insides versus categories. *Behavioral and Brain Sciences*, *37*(5), 502-503.

Sutherland, S. L., & Cimpian, A. (2015). An explanatory heuristic gives rise to the belief that words are well suited for their referents. *Cognition*, *143*, 228-240.

Tworek, C. M., & Cimpian, A. (2016). Why do people tend to infer ought from is? The role of biases in explanation. *Psychological Science*, *27*, 1109-1122.