Seeking Ideal Explanations in a Non-Ideal World

Elizabeth Kon (ellie.kon@berkeley.edu), Tania Lombrozo (lombrozo@berkeley.edu)
Department of Psychology, University of California, Berkeley, 3210 Tolman Hall,
Berkeley, CA 94720 USA

Abstract

Research has found that when children or adults attempt to explain novel observations in the course of learning, they are more likely to discover patterns that support ideal explanations: explanations that are maximally simple and broad. However, not all learning contexts support such explanations. Can explaining facilitate discovery nonetheless? We present a study in which participants were tasked with discovering a rule governing the classification of items, where the items were consistent two non-ideal rules: one correctly classified 66% of cases, the other 83%. We find that when there is no ideal rule to be discovered (i.e., no 100% rule), participants prompted to explain are better than control participants at discovering the best available rule (i.e., the 83% rule). This supports the idea that seeking ideal explanations can be beneficial in a non-ideal world because the pursuit of an ideal explanation can facilitate the discovery of imperfect patterns along the way.

Keywords: explanation; learning; scientific practice

Introduction

Carl Hempel suggested that two human concerns provide the basic motivation for all scientific research (Hempel, 1962): "man's persistent desire to improve his strategic position in the world by means of dependable methods for predicting and…controlling the events that occur in it," and "his deep concern to *know* the world he lives in, and to *explain*, and thus to *understand*, the unending flow of phenomena it presents to him." Hempel isn't alone in highlighting a special role for explanations in science: others identify explanatory theories as the "real payoff" of science (Pitt, 1988).

Explanation is also posited to play a central role within everyday cognition, often in the context of "intuitive" or "folk" theories. For example, Murphy and Medin (1985) suggest that intuitive theories are constituted by mental explanations. Gopnik (2000) suggests that our motivation to seek explanations supports learning by leading us to construct more accurate causal maps of the world.

Why are explanations at the heart of discovery, both in science and in everyday cognition? In this paper we propose that explanations play an important role in learning – even when the environment does not support ideal explanations – thereby advancing Hempel's first motivation for scientific research: the achievement of a better strategic position in the world through better prediction and control. The value of explanation is thus in large part instrumental (Lombrozo, 2011), with the quest for explanations driving theory construction, and the generation of explanations linking theory to application.

Our proposal is motivated by recent work in cognitive psychology on the role of explanation in learning. This work suggests that the process of seeking explanations prompts children and adults to go beyond the obvious in search of broad and simple patterns, thereby facilitating the discovery of such patterns, at least under some conditions (Lombrozo, 2016).

In the present paper we explore an interesting puzzle that arises from this research. On the one hand, prior work suggests that when people engage in explanation, they aim to achieve an explanatory ideal: obtaining explanations that are underwritten by *simple and exceptionless* generalizations. On the other hand, we know that in real scientific practice and in everyday life, such generalizations are rarely to be found. Could it be that searching for ideal explanations is beneficial in part because it facilitates the discovery of real but imperfect generalizations – e.g., those that involve some complexity and exceptions? In other words, is it beneficial to seek ideal explanations even in a non-ideal world? We report an experiment that suggests that it is: prompting learners to explain makes them more likely to discover a "good" explanation, even if it is not an ideal explanation.

The role of explanation in learning

Decades of research reveal that the process of explaining – even to oneself – can have powerful effects on learning (e.g., Fonseca & Chi, 2011; Lombrozo, 2012; Chi et al., 1989). Several psychological processes contribute to this phenomenon. For example, attempting to explain something can help people appreciate what they do not know (Rozenblit & Keil, 2002), make them accommodate new information within the context of their prior beliefs (Chi et al., 1989; Williams & Lombrozo, 2013), and lead them to draw new inferences (Chi, 2000). There is also evidence that when engaged in explanation, both children and adults seek explanations that are *satisfying*, where satisfying explanations are those that account for what is being explained by appeal to broad and simple rules or patterns (Lombrozo, 2016). For example, Williams and Lombrozo (2010, 2013) found that when presented with an array of items belonging to two categories, adults who were prompted to explain why each item belonged to its respective category (e.g., why robot A is a "glorp" and robot B is a "drent") were more likely than those in control conditions to discover a subtle classification rule that accounted for the category membership of all items on the basis of a single feature (see also Walker et al., 2017). This was true whether participants in the control condition were prompted to describe the category exemplars, to think aloud as they studied them, or to simply engage in free study.

What is it about broad and simple patterns that satisfies the demands of explanation? Or conversely, what is it about patterns with exceptions or additional complexity that *fails* to satisfy the demands of explanation? Recent work by Kon and Lombrozo (in prep., 2017) contrasts two possibilities: that explainers favor exceptionless patterns because such patterns maximize predictive power, or that explainers favor exceptionless patterns because such patterns make for more virtuous explanations – that is, for explanations that exhibit the explanatory virtues of simplicity and breadth. To differentiate these alternatives, they created learning tasks in which participants could achieve perfect predictive accuracy on the basis of two salient features of the stimuli (thus achieving breadth at the expense of simplicity), or potentially discover a more subtle pattern that also supported perfect predictive accuracy, and did so on the basis of a single feature (thus achieving both breadth and simplicity, but at the cost of greater cognitive effort). Participants who were prompted to explain were significantly more likely than those in a control condition to discover the more subtle rule. This suggests that the salient, predictively perfect (but less virtuous) alternative was insufficient to satisfy their explanatory drive. This fits well with a familiar observation from science: the most predictive model isn't always the most explanatory. Explanation seems to require something more than successful prediction.

Despite these synergies between our experimental studies and observations about science, an important puzzle remains: scientists rarely succeed in identifying truly exceptionless laws. Especially within the social sciences, generalizations are invariably imperfect and riddled with exceptions. In some domains, accounting for even 75% of the variance in the manifestation of some property is a notable achievement. Could it be that engaging in explanation motivates learners to search for simple, exceptionless patterns, but that in the course of doing so, *they're also more likely to discover other subtle but imperfect regularities that nonetheless constitute an advance*?

Evidence that this could be so comes from Experiment 3 of Kon and Lombrozo (in prep), in which participants were tasked with learning how to determine whether novel creatures eat flies or eat crabs. Half the participants were prompted to write down an explanation for each observation (i.e., for why a particular creature eats flies or crabs), and half (in the control condition) were prompted to write down their thoughts about that observation. The observations were designed to support two possible generalizations. First, participants could learn to predict the diet of all studied examples on the basis of two features of the stimuli, their habitat *and* age, which was a complex but exceptionless pattern. Second, participants could learn to predict the diet of a majority of studied examples (75%) on the basis of a single feature, snout direction, which was a simple rule, but one with exceptions. Kon and Lombrozo found that participants who were prompted to explain were more likely than those in the control condition to discover each of these rules, presumably because they stumbled across them in their search for an ideal explanation: one that was *both* simple and exceptionless. This finding suggests that even if a simple, exceptionless pattern describes some explanatory ideal that is rarely realized, the pursuit of this ideal could spur meaningful discoveries. In the current experiment, we pursued a more systematic test of this possibility.

Experiment

Our experiment investigates whether in the absence of an ideal pattern (i.e., one that is both maximally simple and broad), engaging in explanation can nonetheless facilitate the discovery of the best available alternatives. To test this, we designed a task in which participants learned to classify items into one of two categories. As they studied twelve labeled exemplars (six from each category), they were prompted either to explain or to write down their thoughts about the category membership of the exemplars. Two rules could be used to categorize the items. One rule was fairly salient and therefore easy to discover, but only captured the category membership of 8 of the 12 exemplars (it was thus a "66% rule"). Another rule was much more subtle, but captured the category membership of 10 of the 12 exemplars (it was thus an "83% rule"). So while the latter rule still fell short of the ideal (i.e., a rule that captured all 12 items, a "100% rule"), it was superior to the initial rule along the dimension of breadth.

If explaining facilitates the discovery of the best possible rule, even if it is imperfect, we would expect participants prompted to explain to be more likely than those in the control condition to discover the 83% rule. By contrast, if effects of explanation are restricted to the ideal case – an exceptionless rule – then we would expect participants prompted to explain to perform no better than those in the control condition.

In addition to the *non-ideal world* condition just described, we also considered an *ideal world* condition, in which the more salient rule accounted for 83% of cases. This more familiar situation is a replication of

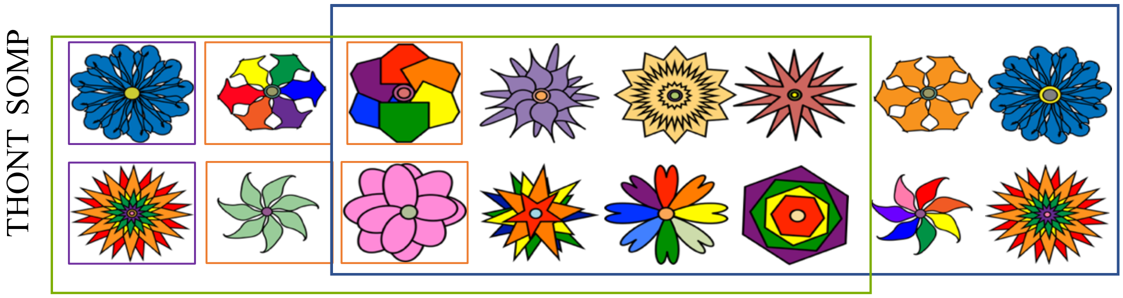

Figure 1: Flower Stimuli. For these flower stimuli, the better rule (100% in the ideal world condition and 83% in the non-ideal world condition) is that SOMP flowers have two concentric circles in their centers whereas THONT flowers have one circle in their center, and the worse rule (83% in the ideal world condition and 66% in the non-ideal world condition) is that the petals of SOMP flowers are mostly one color, while the petals of THONT flowers are mostly rainbow-colored. Within this figure, the green box contains the items in the non-ideal world condition, and the blue box contains the items in the ideal world condition. The exceptions to the worse rule are in orange boxes and the exceptions to the better rule are in purple boxes

prior research, but with a larger number of training exemplars (12 versus 8) to accommodate intermediate percentages. We included this condition in the experiment in part as an extension of prior research, but also to serve as a basis for comparison against the non-ideal world condition. Thus we can ask not only whether a prompt to explain facilitates discovery of a "better" rule when the better rule is an 83% rule (versus a 66% worse rule), but also whether the magnitude of this effect is comparable to the effects of explanation when the "better" rule is a 100% rule (versus an 83% worse rule).

Method

Participants

The sample consisted of 1293 adults[1] (after exclusions)[2] recruited through Amazon Mechanical Turk and paid for their participation. Participation was restricted to adults with an IP address within the United States, and with an approval rating of at least 95% on 50 or more previous tasks. The mean age of participants was 34 (SD = 11, min = 18, max = 81); 509 participants identified as male and 777 as female.

Materials

The stimuli consisted of ten sets of twelve items. The twelve items in each set depicted flowers, containers, objects, simple robots, or complex robots from the *ideal world* or the *non-ideal world*. Throughout this paper we will use flowers as an illustrative example.

Each set contained items from two categories, with six items belonging to each category. For each set, participants could use two possible rules to determine which category an item belonged to. One rule was always "better" in the sense that it could be used to correctly categorize more items than the "worse" rule. In the *ideal world* condition, the better rule was a "100% rule," and the worse rule was an "83% rule" (see Figure 1). For the *non-ideal world* condition, the better rule was an "83% rule," and the worse rule was a "66% rule."

Procedure

The task consisted of a study phase followed by a reporting phase and a rule rating phase. At the start of the study phase, participants were randomly assigned to one of four conditions, which were created by crossing two prompt-types, *Explain* or *Write Thoughts,* with two pattern-types, *ideal world* or *non-ideal world.* Participants were randomly assigned to see one of the stimulus sets.

In the study phase, all participants were told to study the items, and that after the study phase they would be asked questions about how to determine which category each item belongs to. Participants were presented with a randomized array of the twelve items corresponding to their condition's pattern-type *(ideal world* or *non-ideal world).* They were then prompted to focus their attention on each item, individually, in a random order, with a prompt determined by the experimental condition to which they were randomly assigned. Participants in the *explain* conditions were told (for example) to "try to *explain why* flower A is a SOMP flower." Participants in the *write thoughts* conditions were told to *"Write out your thoughts* as you learn to categorize flower A as a SOMP flower." Participants were given 50 seconds to respond to each prompt by typing into a text box, at

[1] Data are from two collections; results are the same within each subsample (see footnote 6 for full details).

[2] An additional 1007 participants failed attention or memory checks (see footnote 4) and were therefore excluded from analyses. We indicate any cases in which these exclusions affect the statistical significance of results.

which time their responses were recorded and the prompt for the next item appeared.

In the reporting phase, participants were asked to report all patterns that they noticed that differentiated SOMPS and THONTS, even if the patterns were imperfect. In addition to describing the rule they discovered in a free-response box, participants were asked how many of the twelve items they thought followed the rule.

After finishing the reporting phase, participants were again presented with all twelve items as well as four candidate rules, presented in a random order, purporting to explain "why flowers A-F are SOMPS (as opposed to THONTS)." We will not consider this rating data here.[3]

Before concluding the experiment, participants completed an attention and memory check question that served as the basis for participant exclusion. This consisted of a fairly long passage that asked them to select "None of these objects look familiar" and to write in the category of the item they recognized. Finally, participants were asked to report their age and sex.

Results

Overall rule reporting. Participants reported finding an average of 1.23 patterns *(SD* = 1.23, min = 0, max = 9) that they reported accounted for an average of 8.18 exemplars *(SD* = 3.13, min = 0, max = 12). Reported patterns were coded for mention of the better and/or worse rule.

Better rule reporting. To test whether explanation prompts affected discovery of the better rule (100% or 83%, depending on pattern-type), and whether effects differed across pattern-type (see Figure 2), we conducted a logistic regression predicting whether participants *discovered the better rule* (yes vs. no) by *prompt-type* (explain vs. write thoughts) x *pattern-type* (ideal world vs. non-ideal world) x *stimulus-type* (flowers vs. containers vs. objects vs. simple robots vs. complex robots). This revealed a significant effect of prompt-type on reporting the better rule (χ^2 = 17.23, p < 0.01), with higher discovery rates for participants prompted to explain. There was also a significant main effect of pattern-type, with more participants reporting the better rule when it accounted for more items (χ^2 = 72.38, p < 0.01). The interaction term between prompt-type and pattern-type was not significant (χ^2 = 0.63, p = 0.43). The interaction term between prompt-type and stimulus-type was also not significant (χ^2 = 6.99, p = 0.14).[4] These findings suggest that explaining indeed facilitated discovery of the better rule, regardless of whether the better rule was ideal, and across a range of different stimulus types.

The results of this analysis are consistent with the hypothesis that when explaining, people seek simple and exceptionless rules, but that in the course of doing so, they are likely to discover "good" rules that may nonetheless fall short of this ideal. To verify this pattern of results for each pattern-type, we ran additional logistic regressions for the ideal world condition and non-ideal world condition separately. We found that explainers reported the better rule significantly more often than those who wrote their thoughts within the ideal world condition (χ^2 = 15.53, p < 0.01) and also within the non-ideal world condition (χ^2 = 3.94, p = 0.05).[5] These results further support the claim that engaging in explanation can facilitate discovery of the best available rule, even when it is imperfect.

Worse rule reporting. Previous studies have found that prompting participants to explain can sometimes decrease worse rule reporting relative to a control condition (e.g., Edwards et al., 2013; Williams & Lombrozo, 2010, 2013). To analyze worse rule reporting we ran another logistic regression: *discovered the worse rule* (yes vs. no) by *prompt-type* (explain vs. write thoughts) x *pattern-type* (ideal world vs. non-ideal world) x *stimulus-type* (flowers vs. containers vs. objects vs. simple robots vs. complex robots). The effect of prompt-type was not significant (χ^2 = 0.39, p = 0.53). The effect of pattern-type was significant (χ^2 = 84.79, p < 0.01): participants reported the 83% worse rule more often than the 66% worse rule. However, the interaction between prompt-type and pattern-type was not

[3] Data on the complex robot stimuli were collected separately from the other 4 stimulus-types and participants did not complete the rule rating phase. We combine the data here because the experimental questions and results were the same.

[4] There was also a significant main effect of stimulus-type (χ^2 = 112.98, p < 0.01), and a significant interaction between pattern-type and stimulus-type (χ^2 = 13.72, p < 0.01).

[5] We initially collected a smaller sample size (1309 participants before exclusions), but the statistical analyses were inconclusive. Specifically, we found the expected effect of explanation (with more participants reporting the better rule when prompted to explain), but we also (a) failed to find an interaction between pattern-type and prompt-type, suggesting that the effects of explanation were comparable across the *ideal* and *non-ideal* world conditions, and (b) failed to find a significant effect of the explanation prompt when restricting analysis to the *non-ideal world* condition, suggesting that explanation did *not* have an effect under these conditions. Because (a) and (b) supported different conclusions, we collected additional data. It is worth noting that while increasing the sample size did change the statistical significance of the effect of explanation within the non-ideal world condition, the proportions of participants reporting the rules remained fairly unchanged by the increased sample size (approximately 15% of the explainers reported the imperfect better rule in both the initial and increased sample, and approximately 10% of control participants reported the imperfect better rule in the initial sample, and approximately 9% reported it in the increased sample). This suggests that the initial sample was simply underpowered.

significant ($\chi^2 = 1.00$, $p = 0.32$).[6] These findings suggest that while explaining improved discovery of the better rule, it did so at no cost to discovery of the worse rule.

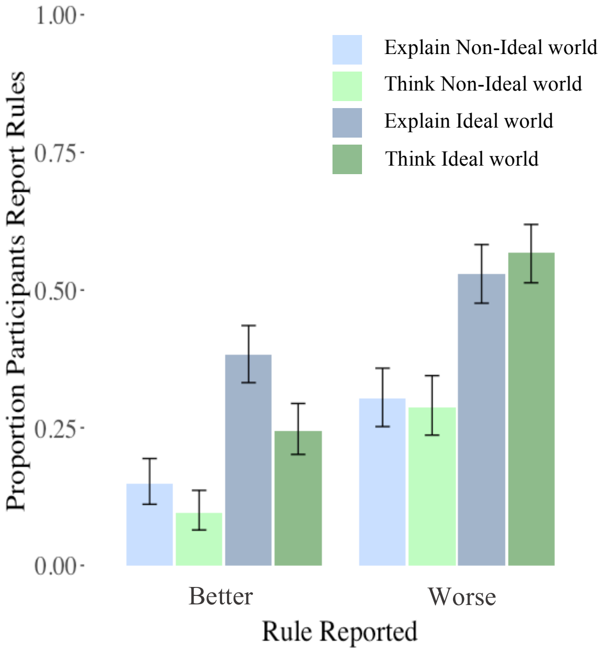

Figure 1: The proportion of participants reporting each rule in Experiment 1, as a function of rule type, condition, and prompt. Error bars correspond to 95% confidence intervals.

Discussion

The results of our experiment both replicate and extend prior research. Consistent with prior research, we found that a prompt to explain facilitated discovery of a subtle, exceptionless rule. Going beyond prior research, we also found some support for the possibility that a prompt to explain facilitates the discovery of a subtle rule that involves exceptions, albeit *fewer* exceptions than a more salient alternative. This helps resolve the puzzle with which we began. On the one hand, scientists and everyday learners are often driven to achieve an explanatory ideal with a prominent role for exceptionless laws and theories that support simple explanations. On the other hand, regularities in the natural world quite often have exceptions, and simple explanations are not always forthcoming. Our findings suggest that the process of seeking ideal explanations may be beneficial because it supports discovery, and that these beneficial effects on discovery are not restricted to the ideal case: explaining can facilitate the discovery of subtle patterns even when those patterns do not account for all cases. This finding is broadly consistent with the idea of "Explaining for the Best Inference" (EBI), introduced by Wilkenfeld and Lombrozo (2015): the process of explaining can sometimes be beneficial because it has positive downstream consequences on what we learn and infer.

Needless to say, our artificial learning tasks are a poor match to many real-world cases, and our classification rules are a poor match to rich scientific or folk-theoretic explanations. The research we present here is no substitute for more ecologically valid approaches and more naturalistic studies of advance within and beyond scientific cases. That said, we expect the learning mechanisms documented here to apply quite broadly. For example, findings concerning the effects of explanation in artificial classification tasks (Williams & Lombrozo, 2010) have been replicated with property-generalization tasks that involve meaningful causal explanations (Kon & Lombrozo, in prep, 2017). The core phenomena found with adults have also been successfully replicated with preschool-aged children (Walker et al., 2014, 2017). These findings suggest that effects of engaging in explanation are fairly widespread and baked into our explanatory activities from a young age.

Although our findings extend the range of contexts in which explanation prompts have been shown to be beneficial, it is worth noting that there are known contexts in which explanation is *not* beneficial. In particular, Williams, Lombrozo, and Rehder (2013) found that if there is *no pattern* available to be found, explanation is actually harmful: participants seem to perseverate in looking for one, leading to erroneous overgeneralizations. Our current findings show that as long as there is *some pattern* to be found that is beneficial in a given learning task, explanation can promote its discovery.

It is also worth noting that there could be interventions other than explanation that lead to the kinds of benefits explanation tends to produce. In particular, if explaining helps by making learners accommodate information in the context of their prior beliefs (Williams and Lombrozo, 2013) and extract the broader statistical structure of a category (Edwards et al., 2013), then we might expect that other interventions that also accomplish this could have a similar effect. We have some evidence that this is the case: in a follow-up to the experiment reported here, participants received an easier version of the task in which items were grouped by category in two 3x2 groups, and participants were asked to study all items of a category together rather than being asked about each individually. This change in presentation format had no effect on explainers, but led to a significant benefit for non-explainers, leading them to explainer-level performance. Studies along these lines can help pin down the mechanisms by which explanation

[6] The effect of stimulus-type was also significant ($\chi^2 = 57.08$, p < 0.01); no interactions were significant (without exclusion criteria, the interaction between pattern-type and stimulus-type was significant ($\chi^2 = 13.79$, p = 0.01)).

generates the observed effects, and which are shared with other cognitive processes.

Some limitations of these studies should be acknowledged. Our participant pool was restricted to online participants within the United States, our learning tasks occurred over a short time scale, and participants were almost certainly more motivated to receive their pay than to uncover the structure of our artificial worlds. Moving forward, it will be important to pursue research that preserves the experimental control of the studies we present here while simultaneously overcoming these limitations. It's also worth noting that some participants in our control condition almost certainly engaged in explanation spontaneously; our comparison is truly between non-prompted explanation and (what we presume to be higher levels of) prompted explanation. If anything, though, this makes the existence and magnitude of our effects more impressive.

Zooming out, our findings support a functionalist approach to explanation (Lombrozo, 2011). On this view, explanation is crucial to science and everyday cognition because it serves an instrumental role. By pursuing explanations of the natural world, we're more likely to generate discoveries and develop theories that in turn improve our strategic position in the world, satisfying Hempel's first motivation for science by pursuing the second.

Acknowledgements

We thank the CoCo Lab. This work was supported by an NSF Career Grant awarded to T.L. (DRL-1056712), as well as a McDonnell Foundation Scholar Award in Understanding Human Cognition.

References

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in instructional psychology*, 5, 161-238.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13, 145-182.

Edwards, B.J., Williams, J.J., & Lombrozo, T. (2013). Effects of explanation and comparison on category learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 406-411). Austin, TX: Cognitive Science Society.

Fonseca, B., & Chi, M. T. H. (2011). Instruction based on self-explanation. In R. Mayer & R Alexander (Eds.), Handbook of research on learning and instruction. New York: Routledge

Gopnik, A. (2000). Explanation as orgasm and the drive for causal understanding: The evolution, function and phenomenology of the theory-formation system. In F. Keil & R. Wilson (Eds.) *Cognition and explanation*. Cambridge, Mass: MIT Press. 299-323.

Hempel, C. (1962). Explanation in science and history. In Colodny, R. G., editor, Frontiers of Science and Philosophy, 7-33. Allen & Unwin, London.

Kon, E., & Lombrozo, T. (in prep). Why explainers take exception to exceptions.

Kon, E. & Lombrozo, T. (2017). Explaining guides learners towards perfect patterns, not perfect prediction. In *Proceedings of the 39th annual conference of the cognitive science society* (pp. 682-687).

Lombrozo, T. (2011). The instrumental value of explanations. Philosophy Compass, 6, 539-551.

Lombrozo, T. (2012). Explanation and Abductive Inference. In K. J. Holyoak & R. G. Morrison (Eds.), The Oxford handbook of thinking and reasoning. Oxford, UK: Oxford University Press.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. Trends in Cognitive Science, 20(10), 748-759.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*. 92(3), 289-316.

Pitt, J. C. (1988). Theories of explanation. Oxford University Press.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science, 26(5), 521-562.

Walker, C. M., Bonawitz, E., & Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. Psychonomic Bulletin and Review. 24(5), 1538-1547.

Walker, C.M., Lombrozo, T., Legare, C., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. Cognition, 133, 343-357.

Wilkenfeld, D. A. & Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). Science and Education, 24(9-10), 1059-1077.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. Cognitive Science, 34(5), 776-806.

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. Cognitive Psychology, 66, 55-84.

Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*. *142*(4), 1006-1014.