Explaining Guides Learners Towards Perfect Patterns, Not Perfect Prediction

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

When learners explain to themselves as they encounter new information, they recruit a suite of processes that influence subsequent learning. One consequence is that learners are more likely to discover exceptionless rules that underlie what they are trying to explain. Here we investigate what it is about exceptionless rules that satisfies the demands of explanation. Are exceptions unwelcome because they lower predictive accuracy, or because they challenge some other explanatory ideal, such as simplicity and breadth? To compare these alternatives, we introduce a causally rich property explanation task in which exceptions to a general rule are either arbitrary or predictable (i.e., exceptions share a common feature that supports a "rule plus exception" structure). If predictive accuracy is sufficient to satisfy the demands of explanation, the introduction of a rule plus exception that supports perfect prediction should block the discovery of a more subtle but exceptionless rule. Across two experiments, we find that effects of explanation go beyond attaining perfect prediction.

Keywords: explanation; learning; causal reasoning

Introduction

"The great tragedy of science - the slaying of a beautiful hypothesis by an ugly fact." T. H. Huxley (1870)

The best explanations account for all the data we invoke them to explain. But in science and in life, explanations often have exceptions. Even when exceptions fail to "slay" our explanatory hypotheses, they certainly diminish them. What is it about exceptions that threatens the quality of explanations?

One possibility is that exceptions are threatening because they offer evidence against the truth of the explanation in question. To the extent our explanation fails to predict an anomalous observation, we might hold out for a better alternative – one that predicts the observation with greater probability, such that the observation provides greater evidential support for that alternative explanation.

A second possibility is that exceptions diminish the quality of explanations not because they reveal predictive failures, but because they reveal that an explanation is deficient with respect to some other explanatory ideal. Across philosophy and science, we praise explanations for their simplicity, breadth, generality, and ability to unify a diverse range of phenomena. Exceptions may diminish the quality of explanations because they threaten these ideals.

In the current experiments, we test these alternatives by investigating how the process of explaining affects learning (for reviews, see Fonseca & Chi, 2011; Lombrozo, 2012). Prior work has found that when learners are prompted to explain, they're more likely to discover regularities that support "good" explanations (Lombrozo, 2016). In particular, Williams and Lombrozo (2010) found that when learning to classify robots from novel categories, those participants who were prompted to explain why each exemplar might belong to its respective category were significantly more likely to discover a subtle classification rule that accounted for all eight items (the 100% rule), as opposed to settling for a more salient classification rule that only accounted for six ("the 75% rule"), leaving two exceptions.

The results of Williams and Lombrozo (2010) support the idea that explaining encourages learners to find an exceptionless pattern, but do not reveal what it is about exceptions that makes the 75% rule less good than the 100% rule. If explaining drives learners away from exceptions because they decrease predictive accuracy, then a rule with non-arbitrary exception – that is, with exceptions that can be reliably identified a priori, such that predictive accuracy can reach 100% – should rival an exceptionless rule. In contrast, if exceptions are undesirable because they threaten some other explanatory virtue, such as simplicity or breadth, then even a rule with non-arbitrary exceptions should be dominated by a 100% rule that classifies all items in a unified way.

To test these predictions, we had participants learn novel relationships while prompted to explain or write down their thoughts, and where the exceptions to the 75% rule were either arbitrary (as in prior work) or meaningful (in the sense that they supported perfect prediction by representing a "rule plus exception" on the basis of two features). If prompting learners to explain pushes them to find a simple, exceptionless pattern, then the two conditions should yield similar results, whether or not the exceptions are meaningful. On the other hand, if explainers are satisfied by a rule that supports perfect prediction, then discovery of the relatively salient 75% rule with meaningful exceptions should block discovery of the more subtle 100% rule. We test these competing predictions in Experiment 1 using a sequential training procedure, and in Experiment 2 using a prediction task.

Our task and stimuli go beyond prior work in a second way, as well. Instead of using a classification task in which participants explain category membership by appeal to arbitrary features, we use a causally-rich property explanation task. Prior work suggests a preference for exceptionless, single-feature rules in classification (e.g., Norenzayan et al., 2002; but see Murphy, Bosch, & Kim, 2016); explanation could simply heighten this classification-based preference. In the current studies, rather than explaining category membership, participants explain why novel creatures eat flies or eat crabs, where both the 75% and 100% rules reflect plausible causal explanations. If prompting learners to explain still promotes discovery of a 100% rule with these modified stimuli, it would suggest that previously-documented effects of explanation on learning are not restricted to classification tasks (see

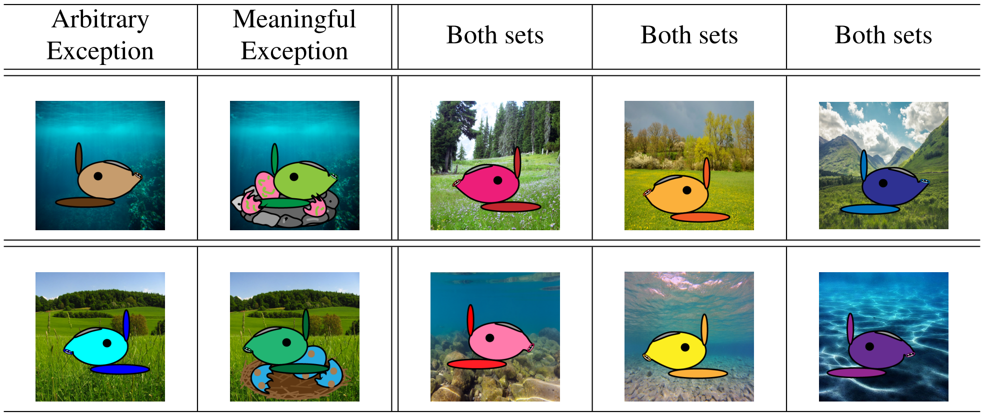

Figure 1: All Stimuli. The top row creatures all eat flies, and the bottom row eat crabs. For the *arbitrary exception* exception-type, participants saw the creatures in the first column, and for the *meaningful exception* exception-type, they saw the creatures in the second column. Both stimulus sets included creatures in columns three through five.

also Walker et al., 2017). This finding would also help bridge the gap between laboratory studies involving artificial materials and educational materials such as biology texts, where effects of explanation have been documented and inform curricula (Fonseca & Chi, 2011).

Experiment 1

Experiment 1 investigates whether engaging in explanation encourages learners to seek simple, exceptionless rules, or to instead find rules that allow for perfect predictive accuracy. To test this, we created two stimulus sets: one with a "meaningful exception rule" (a 75% rule with exceptions identifiable by the presence of a second feature), and another with an "arbitrary exception rule" (a 75% rule with exceptions that do not share a common feature). The meaningful exception rule was relatively easy to discover and supported perfect prediction, but not on the basis of a single feature. If prompting learners to explain makes them persist in seeking an exceptionless single-feature rule, we would predict comparable results for the meaningful exception stimuli and the arbitrary exception stimuli, with learners prompted to explain significantly more likely than those in a control condition to discover the more subtle 100% rule. On the other hand, if perfect predictive accuracy satisfies the demands of explanation, we would expect discovery of the more salient meaningful exception rule to block discovery of the 100% rule, yielding an attenuated effect of explanation on 100% rule discovery, and a boost in discovery of the meaningful exception rule.

Method

Participants Participants were 443 adults recruited from the Amazon Mechanical Turk marketplace. Of these, 124 failed attention or memory checks (described below) or left questions blank and were therefore excluded from analyses. The statistical significance of results are unchanged when these participants are included.

Materials The stimuli consisted of two sets of eight "creatures" each, four of which ate flies and four of which ate crabs (see Figure 1). For each set, participants could use two possible rules to determine whether a creature ate flies or crabs. The first accounted perfectly for all eight creatures (the "100% rule"): all four creatures that ate flies had snouts pointing up; all four creatures that ate crabs had snouts pointing down. The second rule accounted for six of the eight creatures (the "75% rule"): three of four creatures that ate flies were on land; three of four creatures that ate crabs were underwater. Importantly, both features of interest (snout direction and habitat) supported plausible causal explanations for why a creature eats flies versus crabs, e.g., "It eats flies because its snout is pointed up, so it can reach flies" or "It eats flies because it lives on land, where flies are found."

The two stimulus sets differed in the nature of the exceptions to the 75% rule. For participants in the *arbitrary exceptions* condition, the exceptions to the 75% rule did not share a meaningful, plausible characteristic on the basis of which they could be identified as exceptions. For participants in the *meaningful exceptions* condition, the exceptions to the 75% rule were "newborns"– they were green and shown with eggs in a nest. We refer to this manipulation as "exception-type."

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 exception-types, *arbitrary* or *meaningful*.

In the study phase, all participants were told to study the creatures, and that after the study phase they would be asked questions about how to determine which food a creature eats. To provide context and help participants interpret the images, they were told that the creatures were: "from the planet ZARN: the adults of all of these creatures eat either flies or crabs. Newborn creatures look exactly like their adult forms except that they are green because they photosynthesize. There are different subspecies of this animal with different properties. However, they all have a mouth on an inflexible snout, and an ear that sticks up. They are all tailless, born from eggs and have a 4-chambered heart." Participants were presented with a randomized array of the eight creatures corresponding to their condition's exception-type (*arbitrary* or *meaningful*). They were then prompted to focus their attention on each creature, 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 to "try to *explain why* creature X eats flies/crabs." Participants in the *write thoughts* conditions were told to "*Write out your thoughts* as you learn that creature X eats flies/crabs." Participants were given 50 seconds to respond to each prompt, at which time their responses were recorded and the prompt for the next item appeared.

In the reporting phase, participants were told that "we're interested in any patterns that you noticed that might help differentiate creatures that eat flies and creatures that eat crabs. For example, did most or all of the fly-eaters you studied tend to have one property, and most or all of the crab-eaters you studied have another property? We're going to ask you to list all of the patterns (differences between fly-eaters and crab-eaters) that you noticed, one at a time. PLEASE REPORT ANY PATTERNS THAT YOU NOTICED, EVEN IF THEY WEREN'T PERFECT AND EVEN IF YOU DON'T THINK THEY'RE IMPORTANT." This language, adapted from Edwards, Williams, and Lombrozo (2013), was employed to encourage participants to report the 75% rule even if they thought it was incidental or superseded by the 100% rule. In addition to describing the rule they discovered in a free-response box, participants were asked how many of the eight items followed the rule.

After finishing the reporting phase, participants were again presented with all eight creatures as well as four candidate explanations (presented in a random order) for "why creatures A-D eat flies (as opposed to crabs)." They were forced to stay on the page for at least 15 seconds to ensure that they read the explanations (there was no upper time limit). Along with an inaccurate explanation included as a control, the explanations provided for rating were:

- 100% rule: "Because creatures A-D have snouts that point up, and creatures E-H have snouts that point down."
- 75% rule: "Because creatures A-D live on land, and creatures E-H live in the water."
- 75% rule + exception associated with their exception-type:
  - *with arbitrary exceptions*: "Because creatures A-D live on land, and creatures E-H live in the water (with some exceptions)."
  - *with meaningful exceptions*: "Because creatures A-D live on land, and creatures E-H live in the water (with the exception of newly-hatched creatures, who are born in the opposite environment)."

Ratings were collected on a 7-point scale with anchors at 1 ("Very Poor Explanation") and 7 ("Excellent Explanation").

Before concluding the experiment, participants completed an attention and memory check question that served as the basis for participant exclusion. They were asked to "look at the following images and select the one that you have studied in previous questions. In the text box next to that image, please also type in whether you think that it eats flies or crabs. It is important for us to know whether our participants are paying attention and are reading all of the instructions, so if you are reading this, what we actually want you to do is to select "None of these objects look familiar," and in the corresponding text box to write in whether the image you recognize from the other options eats flies or crabs." By selecting the instructed button, participants indicated they had been reading instructions, and by correctly reporting the diet of the creature they recognized, participants indicated that they attended to the stimuli in the primary task.

Results

Overall, participants reported finding an average of 1.25 patterns (SD = 0.96, min = 0, max = 4) that they reported accounted for an average of 5.94 exemplars (SD = 1.8, min = 0, max = 8). Reported patterns were coded for mention of the 100% rule and/or the 75% rule.

100% rule reporting: To test whether explanation prompts affected 100% rule discovery, and whether effects differed across exception-type, we conducted a logistic regression predicting whether participants *discovered 100% rule* (yes vs. no) by *prompt-type* (explain vs. write thoughts) $\times$ *exception-type* (arbitrary vs meaningful). This revealed a significant effect of prompt-type on reporting the 100% rule, collapsed over exception-types (χ^2 = 6.64, p = 0.01; see Figure 2). The interaction term between prompt-type and exception-type was not significant (χ^2 = 0.28, p = 0.6).

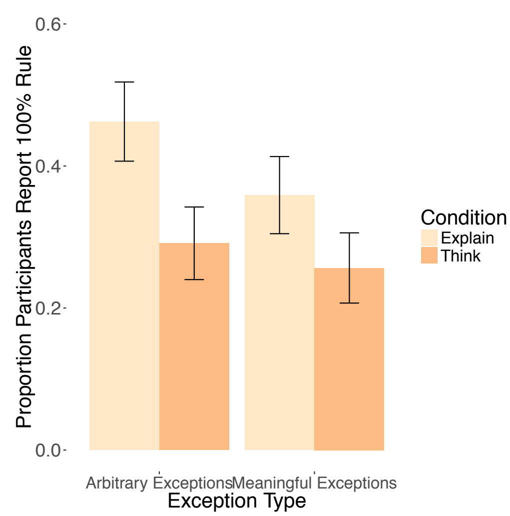

Figure 2: Proportion of Participants Reporting the 100% Rule in Experiment 1

The results of this analysis are consistent with the hypothesis that what people seek when explaining are rules high in explanatory virtues such as simplicity and breadth: the opportunity to employ a rule + meaningful exception (which was both easy to discover and afforded perfect prediction) did not block participants in the *explain* condition from seeking an alternative that accounted for all items with a single feature. However, this conclusion should be accepted with some caution: when analyzed alone, there was not a significant effect of prompt-type within the *meaningful exceptions* conditions (χ^2 = 1.93, p = 0.16), but there was in the *arbitrary exceptions* condition (χ^2 = 5, p = 0.03).

75% rule reporting: Previous studies have found that prompting participants to explain can decrease 75% rule reporting relative to a control condition (e.g., Edwards et al., 2013; Williams & Lombrozo, 2010, 2013). In this study, the proportions of participants reporting the 75% rule were: 51% for explain/arbitrary; 46% for write thoughts/arbitrary; 69% for explain/meaningful; and 63% for write thoughts/meaningful.

To analyze these data we ran another logistic regression: *discovered 75% rule* (yes vs. no) by *prompt-type* (explain vs. write thoughts) × *exception-type* (arbitrary vs. meaningful). The effect of prompt-type was not significant (χ^2 = 1.15, p = 0.28). The effect of exception-type was significant (χ^2 = 10.08, p < 0.01). However, the interaction between prompt-type and exception-type was not significant (χ^2 =0.02, p = 0.9). So while people were more likely to report the 75% rule when the exceptions were meaningful, this effect was not moderated by prompt-type. Few participants reported both the 100% and 75% rules: 16% for explain/arbitrary; 9% for write thoughts/arbitrary; 23% for explain/meaningful; and 17% for write thoughts/meaningful.

Rule Rating: To confirm that the manipulation of exception-type had some effect on perceived explanation quality, we compared explanation ratings for the 75% rule + exception as a function of exception type. Indeed, a t-test revealed higher ratings when the exception was meaningful $t(309)$ = -4.3, p < 0.01 (see Table 1 for all mean ratings).

Table 1: Average Rule Rating by Exception-type

Condition	100% rule	75% rule	75% rule + exception	Bad Rule
Arbitrary Exceptions	5.50(2.23)	3.11 (1.95)	4.87 (2.05)	1.81 (1.54)
Meaningful Exceptions	5.19 (2.4)	3.97 (1.92)	5.80(1.79)	1.55 (1.16)

Discussion

On balance, the results from Experiment 1 support the idea that when it comes to the effects of explanation on learning, an explanation that supports perfect prediction can still be deficient if it fails to account for all observations in a unified way. The experiment also suggests that the original effects reported in Williams and Lombrozo (2010) are not restricted to explicit classification tasks with arbitrary features: we successfully reproduced effects of explanation in a property explanation task where explanations were causally meaningful.

Introducing a rule with meaningful exceptions did have significant effects: participants were more likely to report discovering the 75% rule when the exceptions were meaningful (regardless of prompt), and they evaluated the explanation containing a 75% rule to be more satisfactory when the exceptions were meaningful. However, introducing the 75% rule with meaningful exceptions did not block participants prompted to explain from discovering the 100% rule: they seemed to persevere in looking for an exceptionless, single-feature rule rather than settling for a rule that supported perfect prediction on the basis of multiple features. This conclusion is supported by the significant effect of prompt-type on 100% rule discovery, which was not qualified by a further interaction with exception-type. At the same time, we note that when restricting analysis to the meaningful exceptions condition, the effect of explanation was not significant. The results of Experiment 1 are therefore somewhat inconclusive, and we revisit the contrast between arbitrary and meaningful exceptions in Experiment 2.

Experiment 2

Because the results from Experiment 1 were somewhat inconclusive, we ran a new variant of the task. The task used in Experiment 2 was designed to heighten the value of perfect prediction: rather than receiving labelled exemplars at each step, participants attempt to predict the food that each creature eats, receiving feedback as they proceeded. If explanatory judgments track perfect prediction, then participants prompted to explain in this task should be satisfied with a 75% rule when it involves meaningful exceptions, thereby supporting perfect prediction and blocking or attenuating the effect of explanation on 100% rule discovery.

Method

Participants For this study, 164 adults were recruited from the Amazon Mechanical Turk marketplace. Of these, 61 failed the attention and memory checks described above. We note any cases in which relaxing these exclusion criteria affected conclusions regarding statistical significance.

Materials Stimuli were the same as in Experiment 1.

Procedure This task consisted of a study phase and a reporting phase. As in Experiment 1, participants were randomly assigned to one of four conditions, which were created by crossing two prompt-types, *Explain* or *Write Thoughts,* with two exception-types, *arbitrary* or *meaningful*.

In the study phase, participants were presented with the same introductory text as in Experiment 1. They were then given 5 seconds to look over all eight creatures together before being shown the creatures individually in a random order.

When presented with each of the eight creatures individually, participants were asked to determine whether the creature eats crabs or flies. Based on the accuracy of their response, they were then taken to a screen that said either "CORRECT This item does eat flies/crabs" or "INCORRECT This item eats flies/crabs." They were then given 45 seconds to respond to their condition-specific prompt; either "This creature eats flies/crabs. Try to *explain why* this creature eats flies/crabs." or "This creature eats flies/crabs. *Write down* whatever you are thinking." After cycling through all eight creatures, participants went through them a second time, again in a random order, with 30 seconds to respond.

The reporting phase was identical to that of Experiment 1.

Results

Overall, participants reported finding an average of 0.95 patterns (*SD* = 0.96, min = 0, max = 6) which they reported accounted for an average of 6.35 exemplars (*SD* = 1.53, min = 0, max = 8). Reported patterns were coded for mention of the 100% rule and/or the 75% rule.

100% rule reporting: To analyze 100% rule discovery (see Figure 3), we ran a logistic regression of *discovered 100% rule* (yes vs. no) by *prompt-type* (explain vs. write thoughts) × *exception-type* (with arbitrary exceptions vs. with meaningful exceptions). The interaction between prompt-type and exception-type was not significant (χ^2 = 0.23, p = 0.63).

However, there was a significant effect of explanation (collapsed across the two stimulus sets) (χ^2 = 4.15, p = 0.04)[1]. These findings suggest that the presence of a salient rule that supported perfect prediction in the meaningful exceptions condition was insufficient to block discovery of the 100% rule, and therefore support the proposal that explainers preferentially seek simple, exceptionless patterns, not merely perfect predictability.

Again, to see whether the effect of explanation held within the meaningful exceptions condition, we ran a logistic regression predicting *discovered 100% rule* (yes vs. no) by *prompt-type* (explain vs. write thoughts) using only the results from the meaningful exceptions condition. We found that there was again no significant effect of explanation when restricting analysis in this way, (χ^2 = 2.94, p = 0.09).

75% rule reporting: The proportions of participants reporting the 75% rule were: 36% for explain/arbitrary; 32% for write thoughts/arbitrary; 61% for explain/meaningful; and 59% for write thoughts/meaningful.

To analyze these data we ran a logistic regression predicting *discovered 75% rule* (yes vs. no) by *prompt-type* (explain vs. write thoughts) × *exception-type* (arbitrary vs. meaningful). Again, the effect of prompt-type was not significant (χ^2 = 0.06, p = 0.8), the effect of exception-type was

[1] Without exclusion criteria, (χ^2 = 1.87, p = 0.17)

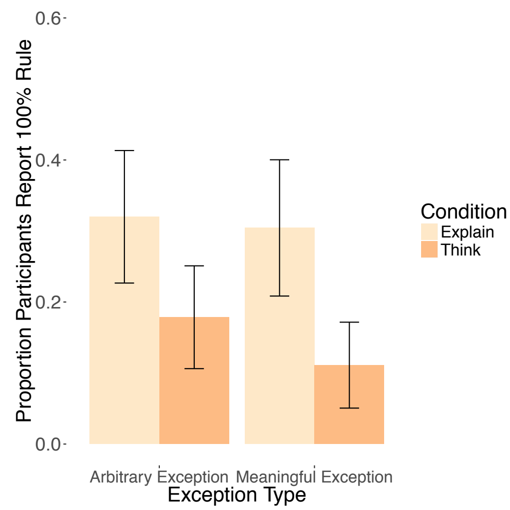

Figure 3: Proportion of Participants Reporting the 100% rule in Experiment 2

significant: (χ^2 = 7.11, p = 0.01), and the interaction between prompt-type and exception-type was not significant (χ^2 = 0.02, p = 0.9). Few participants reported both the 100% and 75% rules: 4% for explain/arbitrary; 0% for write thoughts/arbitrary; 4% for explain/meaningful; and 4% for write thoughts/meaningful.

Prediction Performance: As a check to ensure that the 75% rule with a meaningful exception indeed improved predictability, we additionally analyzed prediction performance in the second block of the task. Specifically, we compared the proportion of exception items that were correctly classified (of 2) as a function of exception-type (arbitrary vs. meaningful) for the 45 participants who reported discovering the 75% rule, but not the 100% rule. A t-test revealed that prediction accuracy was indeed higher when exceptions were meaningful (*M* = 1.39, *SD* = 0.83) than when they were not (*M* = 0.53, *SD* = 0.72), $t(38) = -3.68$, $p < 0.01$.

Discussion

The results of Experiment 2 support the proposal that explainers strive for simple, exceptionless patterns rather than settling for perfect predictability. Even though the presence of meaningful exceptions did improve performance on the prediction task, it did not decrease discovery of the 100% rule differently for participants who explained and for participants who wrote their thoughts.

Discussion

Across two experiments, we find support for the proposal that when explaining, people prefer rules that are high in explanatory virtues (such as simplicity and breadth) over alternative rules that allow for perfect prediction, but that are deficient in these virtues. The threat posed by exceptions therefore appears to be rooted in their disruption of explanatory ideals and not only predictive accuracy. This result is consistent with the observation from science and philosophy that the most predictive models are often not the most explanatory. Additionally, by using a causally-rich property explanation task rather than an arbitrary categorization task, we find support for the claim that effects of explanation on the discovery of exceptionless patterns are not restricted to classification contexts.

Despite these promising results, many questions remain open. First, we found a weaker effect of explanation on 100% rule discovery in the meaningful exceptions conditions than in the arbitrary exceptions condition. This suggests that the presence of a 75% rule that afforded perfect prediction attenuated 100% rule discovery. However, the three-way interaction between 100% rule discovery, prompt type, and exception type did not reach significance, even when pooling results across studies. It thus remains a possibility that introducing meaningful exceptions has a small but real effect on 100% rule discovery; this is worth revisiting with a larger sample and more varied stimuli and learning tasks. Second, our results speak to the consequences of engaging in explanation, but not to the mechanisms by which explaining generates these consequences. The possibility we have advanced is that by virtue of explaining, participants are more likely to reject working hypotheses as they encounter exceptions, and therefore persevere in looking for a pattern that supports a good explanation, where a "good" explanation goes beyond predictive accuracy. Given that participants approach these problems with a host of prior beliefs, future studies should investigate this process more directly, including how learners go about generating hypothesis, seeking information, and updating their beliefs in light of new information.

The fact that explaining can be beneficial in learning is influencing educational systems from online learning environments (e.g. Williams et al., 2014) to college chemistry courses (Teichert & Stacy, 2002). However, as demonstrated here, explanation privileges rules that are simple and exceptionless, and not all learning contexts involve this kind of structure. In fact, previous work has found that prompting learners to explain is sometimes detrimental (e.g. Berthold et. al., 2011; Kuhn & Katz, 2009; Rittle-Johnson & Loehr, 2016; Williams & Lombrozo, 2013; see also Nokes et al., 2011). This underscores the importance of understanding when and why engaging in explanation will and will not promote particular learning outcomes; our current findings provide an additional step towards achieving this understanding.

Acknowledgements

We thank the CoCo Lab, particularly Daniel Wilkenfeld. 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

Berthold, K., Röder, H., Knörzer, D., Kessler, W., & Renkl, A. (2011). The double-edged effects of explanation prompts. *Computers in Human Behavior*, *27*(1), 69–75.

Edwards, B. J., Williams, J. J., & Lombrozo, T. (2013). Effects of explanation and comparison on category learning. In *Proceedings of the 35th annual conference of the cognitive science society* (pp. 406–411).

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

Huxley, T. H. (1870). Biogenesis and abiogenesis. *Collected Essays of Thomas H. Huxley*, *8*, 256.

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology*, *103*(3), 386–394.

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 Sciences*, *20*(10), 748–759.

Murphy, G. L., Bosch, D. A., & Kim, S. (2016). Do Americans Have a Preference for Rule-Based Classification? *Cognitive Science*.

Nokes, T. J., Hausmann, R. G. M., VanLehn, K., & Gershman, S. (2011). Testing the instructional fit hypothesis: The case of self-explanation prompts. *Instructional Science*, *39*(5), 645–666.

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, *26*(5), 653–684.

Rittle-Johnson, B., & Loehr, A. M. (2016). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic Bulletin & Review*, 1–10.

Teichert, M. A., & Stacy, A. M. (2002). Promoting understanding of chemical bonding and spontaneity through student explanation and integration of ideas. *Journal of Research in Science Teaching*, *39*(6), 464–496.

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining Constrains Causal Learning in Childhood. *Child Development*, *88*(1), 229–246.

Williams, J. J., Kovacs, G., Walker, C. M., Maldonado, S., & Lombrozo, T. (2014). Learning online via prompts to explain. In *Chi'14 extended abstracts on human factors in computing systems* (pp. 2269–2274).

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.