How do instructions, examples, and testing shape task representations?

Aba Szollosi[1] (aba.szollosi@gmail.com), **Vlad Grigoraș**[1], **Tadeg Quillien**[2], **Chris Lucas**[2], **Neil Bramley**[1]

[1]Department of Psychology, University of Edinburgh
[2]School of Informatics, University of Edinburgh

Abstract

People need to generate and test hypotheses in order to create accurate representations of their environments. But how do they know which hypotheses to consider when there are often infinitely many possibilities? Here we explore the idea that evolutionary mental representation generation and selection processes – responsible for the generation of both local (i.e., within a task) and global (i.e., about a task) representations – enable people to address this problem. We investigated this through an active learning experiment, where participants' task was to discover a hidden rule determining the behavior of a simple physical system. Specifically, we aimed to manipulate factors that constrain this process, particularly through experimental instructions and feedback. We found that providing more opportunities for participants to recognize when their initial task conceptualization was wrong and adjust it helped them create more accurate representations about the task, which in turn led to better accuracy within the task.

Keywords: active learning; constructive models; hypothesis generation and selection

Introduction

Building and maintaining a representation of the world seems to involve ongoing processes of generating, testing, and revising hypotheses at different levels and scales of abstraction (Szollosi & Newell, 2020). Psychologists have typically investigated human learning by studying behavior in tasks with carefully selected fixed and finite dimensions. However, there is considerable variation in how much effort is then made to bring participants' task conceptualization in line with the experimenters'. We explore the idea that participants often represent tasks quite differently to experimenters, in spite of instruction, unless they are provided with ample opportunities to recognize and correct inconsistencies. We suggest this leads to a pervasive pattern of analyses that mistake representational differences for simple lapses or randomness.

To generate hypotheses worth entertaining (out of an often-infinite set of possibilities) in a particular environment, people first need to create a reasonably good representation of the rules and affordances of that environment. A person trying to pick their next move in a game of chess must first establish how the pieces are allowed to move, what the exceptions to those rules are, and what the goal of the game is. A plumber trying to find the cause of a blockage needs to first have a good model of how pipes work and what stops them from working. A gardener trying to decide which tree to plant in a particular garden needs to first have a good model of the conditions of the garden, and about the kinds of trees that prefer those conditions. The common theme across these situations is that, in order to be able to create and test appropriate hypotheses, people need to have an apt representation of the problem situation in the first place – otherwise they are liable to waste time and resources trying things that will not work or cannot be right.

Despite the importance of developing an accurate representation of the problem situation at hand, theories of learning and decision making often neglect to explain how people achieve this. Instead, they tend to start at the point where the task representation is already built – for example, by assuming that participants have perfectly recreated the intended structure and affordances of an experiment based on its description in the instructions. Yet there is ample evidence that people make errors when developing task representations for a variety of reasons – such as not sharing the background knowledge on which the task is based (Szollosi et al., 2023), inferring not (necessarily) intended implications of the experimenter's communicative acts (Hilton, 1995), or simply not caring enough about the task to invest a lot of mental effort into developing a good representation of it (Tversky & Kahneman, 1974).

This paper aims to start filling this gap in our understanding by exploring the processes by which task representations develop. We construe this as a more general evolutionary-like representation generation and selection process (labelled as such by Campbell, 1960; but also frequently related to "constructivism" in other parts of the literature, e.g., Carey, 2009), according to which all types of representations are generated by alternating cycles of variance-increasing and selective-retention processes (i.e., two alternating processes, the first of which expands or explores the space of possible representations, and the second of which selectively retains some subset of the representations under consideration). These processes enable the creation of not only proximal hypotheses (i.e., within the task) but also more global ones (i.e., about the task itself). They are similar in some ways to the processes of variation and selective retention (through genetic mutation and natural selection) that drive biological evolution.

Although such evolutionary-type theories do not make predictions about modal responses (in contrast with conventional theories in psychology), they make predictions about variability (Campbell, 1960; Szollosi et al., 2023). Specifically, they predict variability in responses (outputs, products, or behaviors) to change as a function of selection pressures: When selection pressures increase, we should expect the range of responses to reduce, and conversely when

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

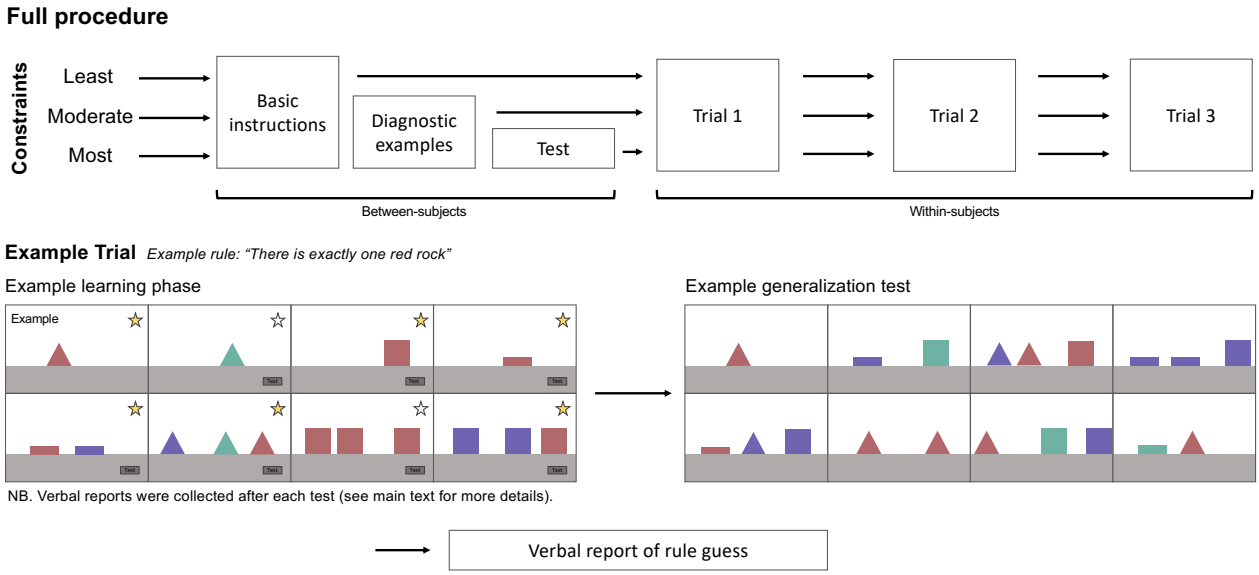

Figure 1: Schematic illustration of the design. *Top panel:* General structure of the task. *Bottom panel:* Example of a trial (actual data of Participant #21).

they decrease, we should expect the range of responses to increase (similar to how historically speciation has arisen predominantly in periods of abundance but decreased during periods of hardship or scarcity). In learning, we can think of selection pressures (or constraints) coming from the act of comparing hypotheses to observations (evidence or feedback from the environment), but they can come equally from other information (such as task instructions, pre-existing representations of similar contexts, etc.). Here, we aimed to probe both sources.

Active learning tasks (e.g., Bramley et al., 2015; Markant & Gureckis, 2014) provide a natural testing ground for such predictions, because they are deliberately unconstrained, allowing a broader view of the hypothesis generation and selection process. Here we used a simple rule-discovery task (based on Bramley et al., 2018) in which participants had to identify a hidden rule guiding the behavior of simple physical environments by actively manipulating and testing those environments. This task models well the relevant characteristics of real environments (i.e., a mixture of known and unknown processes that people need to understand to increase control over the environment – similar to our real-world examples above) and so enables the study of the processes responsible for creating representations of them.

To better understand how people's task representations develop and how this affects their within-task behavior, we aimed to experimentally manipulate constraints on task understanding (through instruction, examples, and comprehension tests *before* the task) and rule sparsity (through environmental feedback *within* the task). We expected that increasing *instructional constraints* – going from only a minimal description of the task, to providing illustrative examples about relevant hypotheses, to testing participants knowledge on what the rules can be – would help participants in narrowing their task representation (i.e., align it better with the intended representation) and consequently in generating more appropriate hypotheses to test. We also expected that by introducing increasingly *constraining rules* (through manipulating rule sparsity, the proportion of possible tests resulting in a positive outcome) would further constrain the set of hypotheses participants entertained – on the assumption that tests with negative outcomes provide more opportunities to learn about the constraints of the space.

Methods

Participants and materials

We recruited 101 participants on Prolific academic. They were paid a flat fee of GBP 2.50. The experiment was based on Bramley et al. (2018) and took participants 25 minutes to complete on average.

General structure of the task

Participants were given an "alien planet" cover story in which their task was to discover a rule for how to arrange rocks so that they emit radiation when heated. The rocks varied along two features: shape (stick, square, triangle) and color (red, green, purple). The rule set was restricted such that they specified that an arrangement must contain either exactly one, two, or three rocks of a specific feature to produce radiation (e.g., "There is exactly one red rock"). A schematic illustration of the procedure can be seen in the top panel of Figure 1.

After receiving instructions imposing varying levels of constraints (manipulated between-subjects, see below),

participants completed three trials each consisting of a learning phase and a generalization phase. Participants had to discover a different rule in each of the three trials (manipulated within-subjects, see below).

In the *learning phase*, participants were shown an initial example arrangement that followed the rule[1] and then performed 7 tests themselves by constructing arrangements on their own (bottom-left panel on Figure 1). Each test scene started with no rocks in it. They were allowed to add and remove any rock to a test arrangement, which could consist of up to three rocks (but at least one). Once they were satisfied with the test arrangement they clicked "Test", which resulted in feedback as to whether the arrangement caused radiation[2].

In the *generalization phase*, participants were asked to identify the rule-consistent arrangements out of 8 possibilities (bottom-right panel on Figure 1). They had to select at least one but could not select all of them. The arrangements were selected in advance such that 4 were rule consistent and the other 4 were not. They were presented in a randomized order to each participant. After this task, participants were asked to verbally state what they thought the rule was ("Now that you had the opportunity to test various arrangements, what do you think the rule is for how to set up arrangements to create radiation on this planet? Try to make your description as clear and specific as possible.").

Experimental manipulations

The experiment used a 3 × 3 design (instruction constraint × rule constraint). For the instruction constraint manipulation, participants received the instructions of varying complexity. In the *least restrictive* condition (n = 37), participants were only informed about the main goal of the experiment (i.e., to discover the hidden rule that produces radiation), and the mechanisms by which they can construct and test arrangements. They had to correctly complete a generic attention check questionnaire (which aimed to test whether they paid attention to surface-level features of the instructions, such as the color and shape of the rocks, and the method to test whether an arrangement causes radiation) to continue and had to repeat the instruction phase if they failed. In the *moderately restrictive* condition (n = 34), in addition to the previous, participants were told that "As you just saw, the rules specify that an exact number of rock(s) of either a specific color or shape must be present for a type of radiation to appear" and were shown example positive and negative tests for three hypothetical rules[3]. In the *most restrictive* condition (n = 30), participants had to additionally categorize five rules as to whether they were possible or impossible under the instructions (e.g., "There are exactly two green rocks" – possible; "There is exactly one purple square-shaped rock" – impossible). They had to repeat the instruction phase if they failed. The aim of this manipulation was to increasingly constrain the ways in which participants could construe the task itself, and therefore the possible hypotheses they might generate and test in the trials.

For the rule constraint manipulation, on each trial, participants were assigned a rule that they needed to discover, which specified the number and feature of rocks that must be present (i.e., that there must be exactly one, two, or three rocks of a specific feature present). Every aspect of this rule was completely hidden from the participants. Each participant received (in a randomized order) a trial where the rule referred to one rock, a trial where the rule referred to two rocks, and a trial where the rule referred to three rocks. The specific feature that the rules referred to (a specific shape or color) was randomly drawn from all possible features on each trial. The aim of this manipulation was to constrain the set of possible arrangements (or, in other words, rule sparsity) that would produce a positive outcome (38.36%, 19.18%, and 4.57% of all possible arrangements for the one, two, and three rule conditions respectively).

Dependent variables

We aimed to assess the effects of our manipulations on multiple measures. Our main dependent variable was how well participants identified the rule as measured by generalization accuracy (i.e., the number of arrangements correctly categorized as rule following or not rule following out of a total of 8 arrangements shown in the test phase). This measure was supplemented by verbal reports of the rule (both their accuracy and variability; see Results for details).

To get at the processes underlying participants' representation building in even greater detail, we also analyzed measures of active learning. We used the difference between the proportion of tests producing positive outcomes and the proportion of such tests expected under random testing as a crude measure of the sensibility of participants' hypotheses (i.e., whether their hypotheses mapped onto some aspects of the real rules at all). To get a better measure of how informative participants' tests were over the course of each trial, we compared them to ideal observer models with initial hypothesis spaces of varying complexity.

[1] This initial example was chosen such that it showed a rock / multiple rocks with the feature (i.e., shape or color) designated by the rule for that trial (more details about rules are given under the Experimental Manipulations section). The other feature(s) of the rock(s) (i.e., color or shape) was/were counterbalanced. No other rocks were shown in these examples.

[2] After each test, participants were asked to give free text response to the question: "Please explain what you expected to learn about the rule by constructing the arrangement in this particular way and why?" This question was included to inspire a follow-up experiment, so we do not analyze these responses in the present paper.

[3] Since the test cases were independently counterbalanced, occasionally they would coincide with one or more of the examples.

Results

We analyzed the data in R (R Core team, 2022) using RStudio (Posit Team, 2022). We used the lme4 package (Bates et al., 2015) for mixed-effect modeling, and the ggplot2 package (Wickham, 2016) for creating the figures.

To evaluate the effects of our manipulations, we predicted the various outcome measures using general or generalized linear models. We added predictors (instruction condition, rule condition, and their interaction) stepwise to the model and compared them to each other in most cases (deviations from this are noted in the relevant sections). Participants had random intercepts in all models. Reported test statistics reflect model comparisons between the two best fitting models.

The rule guesses were categorized by the first two authors as correct or incorrect. Agreement was 85.48%; Cohen's Kappa also indicated substantial agreement $\kappa = .70$, $p < .001$. In cases of disagreement, the first author revisited the rule and made the final decision. In addition, to have a better understanding of the variability of inaccurate rule guesses, the second author developed and applied a categorization scheme for all free text responses based on logical equivalence for the possible scenes in the task.

Accuracy

Figure 2 displays participants' generalization accuracy broken down by experimental conditions. Generalization accuracy was best predicted in the model that included instruction and rule condition, but not their interaction, $\chi^2(2, N = 7) = 53.52$, $p < .001$. Generalization accuracy (i.e., number of arrangements correctly categorized as rule-following/not-rule-following out of 8) increased both as a instructional constraints ($M_{Least} = 5.68$, $SE_{Least} = 0.14$, $M_{Moderate} = 6.32$, $SE_{Moderate} = 0.18$, $M_{Most} = 7.14$, $SE_{Most} = 0.16$) and rule constraints ($M_{One\ X} = 5.84$, $SE_{One\ X} = 0.15$, $M_{Two\ X} = 6.10$, $SE_{Two\ X} = 0.14$, $M_{Three\ X} = 7.05$, $SE_{Three\ X} = 0.14$) increased.

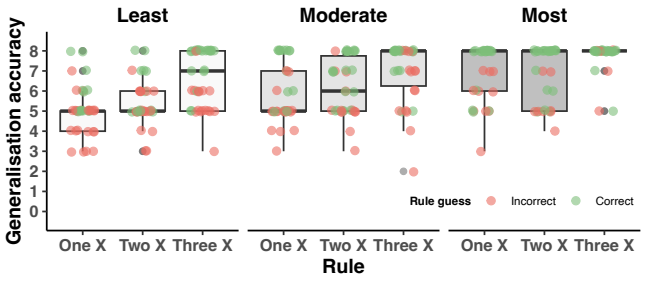

Figure 2: Participants' generalization accuracy (boxplots) and rule-guess accuracy (individual dots; green = accurate, red = inaccurate) broken down by experimental conditions. Individual dots reflect individual participants.

The pattern was the same for verbal reports of the rule (also shown on Figure 2). Accuracy of verbal reports was predicted best by a model that included instruction and rule condition, but not their interaction, $\chi^2(2, N = 6) = 7.39$, $p = .002$. Accuracy improved with increasing instructional constraints and was 44.14%, 52.94%, and 74.44% in the least, moderately, and most restrictive conditions respectively. Similarly, accuracy improved with increasing rule constraints and was 47.52%, 56.44%, and 64.36% in the conditions where rules referred to one, two, or three features respectively.

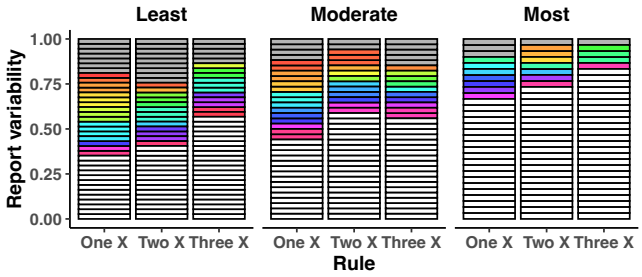

Figure 3: Variability in rule reports. White indicates correctly guessed rule. Grey indicates inaccurate uncategorized guesses. Other colors indicate alternative inaccurate guesses.

Rule guess accuracy also predicted generalization accuracy (compared to a null model), $\chi^2(1, N = 4) = 111.25$, $p < .001$, such that generalization accuracy was higher when participants' guess about the hidden rule was correct, $\beta = 1.73$, $SE = 0.15$.

Figure 3 indicates that guessed rules were increasingly less varied as instructional constraints increased: there were 35, 30, and 19 unique categories of rule reports in the least, moderately, and most restrictive conditions respectively. As an illustration, in Table 1, we provide some examples of participants' guesses.

Table 1: Examples of verbal report guesses.

Guess category (Instruction condition)	Example
Accurate guess ('Most')	"Exactly 3 triangle rocks"
Conjunctive rule ('Moderate')	"2 green square rocks"
Disjunctive rule ('Moderate')	"I think it must include a red triangle, or it could include any coloured triangle"
Uncategorized ('Least')	"Must contain at least 2 squares either both red or 1 red and 1 green"

Active learning

Figure 4 summarizes aggregate positive outcome proportion (i.e., the proportion of arrangements participants generated that produced radiation) across conditions. We expected participants to construct fewer arrangements leading to positive outcomes as rule sparsity increased, since there were fewer arrangements that would yield a positive outcome in those cases (this was because of the restriction on the number of objects that may be added to an arrangement). True proportions were 38.36%, 19.18%, and 4.57% in the

conditions where rules referred to one, two, or three features respectively (indicated by red stars on Figure 4). A model that included rule constraints as a predictor performed better than a null model, $\chi^2(2, N = 4) = 15.27$, $p < .001$, and adding other predictors did not significantly increase model fit. The proportion of tests with positive outcomes was 62%, 57.3%, and 52.6% in the conditions where rules referred to one, two, or three features respectively – substantially higher than what these proportions would have been if arrangements were constructed at random.

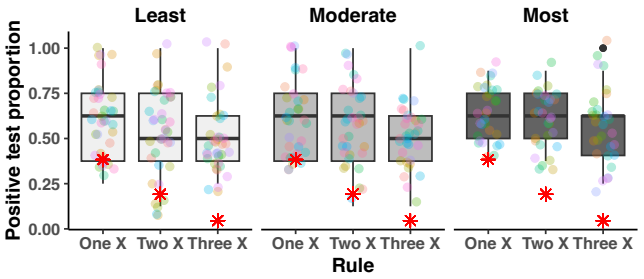

Figure 4: Proportion of tests with positive outcomes broken down by experimental conditions. Individual dots reflect aggregate positive tests for individual participants for each trial. Red stars indicate overall proportion of possible arrangements consistent with the rule.

For a more detailed analysis of the temporal trajectory of participants' active learning, we calculated the Shannon entropy (Shannon, 1951), $H(.)$, on each of their tests with respect to the intended hypothesis space R (i.e., the 18 hypotheses specifying the exact number of a single feature that needs to be present in an arrangement), and also with respect to a larger hypothesis space R^+ (a total of 1118 hypotheses) containing the intended one, plus conjunctive hypotheses (i.e., hypotheses specifying the exact number of conjoined shape and color features that need to be present), and relaxed variations of both (i.e., the number of features could be specified as 'at least' and 'at most' – an example hypothesis from the most relaxed space is "There are at least two blues and there is exactly one blue triangle"). Through this measure, we aimed to quantify the quality of participants' test arrangements based on how much reduction in uncertainty they achieved – taking into account the relevant set of hypotheses, past evidence, and possible other test arrangements. We assumed a uniform prior over legal hypotheses in both spaces. $H(.)$ was calculated according to

$$H(R|D) = -\sum_{r \in R} P(R|D) \log_2 P(R|D) \qquad (1)$$

where $r \in R$ are the hypothetical rules, and D is the data the participant has seen up to that test.

Figure 5 displays the temporal trajectory of entropy for both the intended and extended hypothesis spaces (reflecting the extent to which participants reduced their uncertainty by that point within the respective spaces). For the intended set, only in the clearest instruction condition could all participants reduce their uncertainty to zero (and only for the one/two feature rule conditions) over the course of the 7 tests (for the 1st position we plotted, the example was given to the participants). For the extended set, participants achieved substantial uncertainty reduction in all conditions, although participants in the clearest instruction condition seemed to retain some advantage over other conditions.

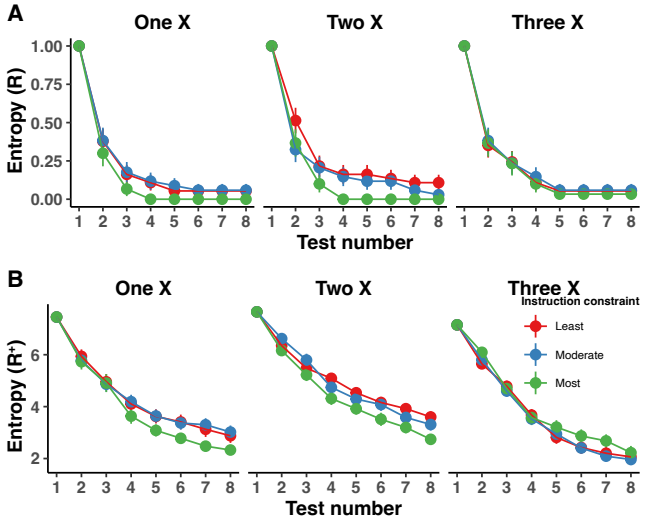

Figure 5: Average change in entropy over tests broken down by experimental conditions. Error bars reflect standard error. A) Change in entropy for the intended hypothesis space. B) Change in entropy for the extended hypothesis space.

Discussion

We investigated how people generate and refine hypotheses in and about an active learning task. We gradually increased constraints through experimental instructions and task feedback. Our results were in line with our expectations: With more constraints imposed, participants' representational variability decreased, and their accuracy increased. We explained this as a result of their task representation becoming better aligned with the intended representation of the task. Although these findings may seem self-evident from a commonsense perspective, they are surprising under accounts of learning and decision making that implicitly assume that people spontaneously represent tasks as intended by the researcher – under such accounts, we should have observed consistently close to ceiling performance.

We found that increasing instructional constraints resulted in both better generalization accuracy and more correct rule guesses. In addition, their final rule guesses were less variable. Participants also appeared to generate more informative tests (and quicker) in the most constraining condition according to normative measures. These results are consistent with our explanation that participants used instructional information to refine their task representation – which led to a narrower (and more appropriate) hypothesis

set to test, which in turn gave them a better chance of discovering the intended rule. The results also dovetail with Bonawitz and colleagues' (2011) finding that children's learning was improved (but their spontaneous exploration decreased) as instructions about how to use a toy were made increasingly clear.

Participants' responses were affected similarly by the constraining influence of feedback caused by the true rule's sparsity. Specifically, in conditions where the expected positive outcome rate was lowest (where all three objects had to have a feature in common), participants' accuracy was highest. Our explanation for this is that because in these conditions participants' hypotheses were exposed to more disconfirmation (i.e., their tests had more negative outcomes), they had more opportunities to learn when they were wrong and could adjust their hypotheses accordingly. Alternatively, however, it could be the case that initially provided examples helped people in narrowing the hypothesis space even before any testing took place, because they could be used as anchors to base further hypotheses on (these anchors could be helpful in figuring out the intended hypotheses because the examples given to the participants did not contain potentially misleading additional rocks, e.g., as shown on Figure 1). Future research may manipulate initial examples independently to tease apart these explanations.

These results further highlight the problem with the general assumption that participants' task representation largely aligns with the experimenters'. The more general evolutionary/constructive view we have advocated here implies that solving a task also involves creating a representation of the task – and if the experimenter does not provide enough help for the participant to recreate the intended representation, then they might end up solving a different task. The misrepresented task might be more difficult to solve (as was the case in some conditions of the current study) and correct answers may differ from those of the intended task – making the evaluation of the adequacy of participants' actions and responses difficult at best. For instance, a researcher may mistakenly conclude that participants are generally bad at solving such tasks, when in fact it was the researcher who did not give instructions appropriately (McKenzie, 2003; Szollosi & Newell, 2020).

A more in-depth consideration of the issue of creating task representations might also be relevant for artificial intelligence research, where such representations are often hand-coded into models (e.g., the rules of chess). If the aim is to make such models more human-like in terms of generality – for example, to allow them to perform better in one-shot learning and/or transfer to novel environments – a better understanding of the processes responsible for domain-general representation generation is likely to be key.

Although the current findings showed how introducing constraints at different points of the representation-generation process influences people's representations about and within a task, it is yet unclear how these constraints get incorporated into the variation and selection learning cycles. One approach is to use constraints just for selection, perhaps discarding and regenerating representations that mismatch feedback. Another approach could be to use feedback to adapt one's construction process such that it is more likely to create appropriate representations (cf. Cropper, 2022). We speculate human learning makes use of both constraint mechanisms. Using grammar-based algorithms, recent work has begun to model learning on such tasks assuming that people continually make local edits to existing representations and preferentially retain those that are better adapted to their objective (e.g., Bramley et al., 2017; Zhao et al., 2022). Such algorithms might also be adapted to model the constraint mechanisms explored in the current paper.

Future research can further investigate implications of the evolutionary-type representation-generation processes we argued for in this paper – for example, identifying where they diverge from probability-respecting mutation algorithms like MCMC, and by probing other "pinch points" in the representation generation process. Here we focused on constraints at the beginning and end of the process (instructions and environmental feedback respectively) – but there are other possibilities, including constraining the process through the information provided in examples, or through teaching people more or less suitable active learning strategies for testing their hypotheses.

Acknowledgments

This research was funded by an EPSRC New Investigator grant (EP/T033967/1) and a Post-Doctoral Enrichment Award by The Alan Turing Institute. We thank the anonymous reviewers for helpful feedback and suggestions.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322-330.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, *124*(3), 301-338.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708-731.

Bramley, N., Rothe, A., Tenenbaum, J., Xu, F., & Gureckis, T. (2018). Grounding compositional hypothesis generation in specific instances. In *Proceedings of the 40th annual conference of the cognitive science society*.

Campbell, D. T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological review*, *67*(6), 380-400.

Carey, S. (2009). *The origin of concepts: Oxford series in cognitive development.* Oxford University Press.

Cropper, A. (2022). Learning Logic Programs Though Divide, Constrain, and Conquer. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(6), 6446-6453.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological bulletin*, *118*(2), 248-271.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General, 143*(1), 94-122.

McKenzie, C. R. (2003). Rational models as theories–not standards–of behavior. *Trends in Cognitive Sciences*, *7*(9), 403-406.

Posit Team (2022). *RStudio: Integrated Development Environment for R.* http://www.posit.co/

R Core Team (2022). *R: A language and environment for statistical computing.* https://www.R-project.org

Shannon, C. E. (1951). Prediction and entropy of printed english. *The Bell System Technical Journal, 30*, 50–64.

Szollosi, A., Donkin, C., & Newell, B. R. (2023). Toward nonprobabilistic explanations of learning and decision-making. *Psychological Review, 130*(2), 546–568.

Szollosi, A., & Newell, B. R. (2020). People as intuitive scientists: Reconsidering statistical explanations of decision making. *Trends in Cognitive Sciences*, *24*(12), 1008-1018.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124-1131.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag. https://ggplot2.tidyverse.org

Zhao, B., Lucas, C. G., & Bramley, N. R. (2022). How do people generalize causal relations over objects? A non-parametric Bayesian account. *Computational Brain & Behavior, 5*(1), 22–44.