Exploring Causal Overhypotheses in Active Learning

Chentian Jiang (chentian.jiang@ed.ac.uk)
School of Informatics, 10 Crichton Street
Edinburgh, EH8 9AB, UK

Christopher G. Lucas (clucas2@inf.ed.ac.uk)
School of Informatics, 10 Crichton Street
Edinburgh, EH8 9AB, UK

Abstract

People's active interventions play a key role in causal learning. Past studies have tended to focus on how interventions help people learn relationships where causes are independently sufficient to produce an effect. In reality, however, people can learn different rules governing how multiple causes combine to produce an effect, i.e., different functional forms. These forms are examples of causal overhypotheses—abstract beliefs about causal relationships that are acquired in one situation and transferred to another. Here we present an active "blicket" experiment to study whether and how people learn overhypotheses in an active setting. Our results showed participants can learn disjunctive and conjunctive overhypotheses through active training, as measured in a new disjunctive task. Furthermore, intervening on two objects led to better conjunctive judgments, and complementarily, conjunctive training predicted more objects in future interventions. Overall, these results expand our understanding of how active learning can facilitate causal inference.

Keywords: causal learning; active learning; transfer learning; overhypothesis; intervention

When learning causal relationships, people often play the role of an active experimenter, choosing actions to intervene on a causal system and observing the consequences of those interventions. Several studies have shown that our interventions tend to be informative, providing more information than passive observations or arbitrary intervention choices (Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis, 2015; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Cook, Goodman, & Schulz, 2011; Sim & Xu, 2017; Oaksford & Chater, 1994). Thus, interventions play a key role in learning and making accurate inferences about causal relationships.

However, past active learning studies focused mainly on discovering the *structure* of a causal relationship—which objects or events are causes and which are their effects (Figure 1). The nature of the relationship, on the other hand, was either explicitly described or consistent with the simple expectation that causes are independently sufficient to produce or prevent an effect. The independence assumption holds for a wide variety of phenomena in causal inference and appears to be a default expectation people have in unfamiliar contexts (Cheng, 1997; Gopnik & Sobel, 2000; Griffiths & Tenenbaum, 2005; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008). In reality, however, this assumption is not always appropriate and people may not know what kinds of causal relationships are plausible in a novel context. For example, a child might have to learn to use multiple simultaneous batteries in a circuit to activate an LED. In such a situation, people must learn different kinds of

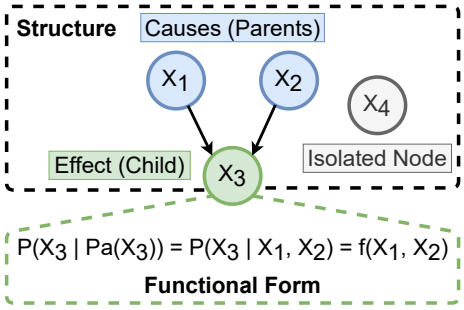

Figure 1: Causal graph. The edges (arrows) define the causal *structure*: For a given edge, the source node is the "parent" (cause) and the destination node is the "child" (effect). Isolated nodes are not involved in any causal relationships. The child node X_3 has a conditional probability P defined by the *functional form* (one type of overhypothesis) of the relationship with its parents $Pa(X_3)$.

causal relationships, e.g., that multiple causes are required to produce an effect, and use these beliefs to guide their causal inferences in new contexts (Lucas & Griffiths, 2010; Griffiths & Tenenbaum, 2009; Griffiths, Sobel, Tenenbaum, & Gopnik, 2011; Lu, Rojas, Beckers, & Yuille, 2016).

Formally, people can learn different *functional forms* of causal relationships—rules governing how multiple causes combine to produce an effect (Figure 1). Functional forms are essential to causal learning because they are examples of *causal overhypotheses*—abstract beliefs about causal relationships that transfer across different contexts, accounting for the prior knowledge that people use to constrain their learning in new contexts (Lucas, Gopnik, & Griffiths, 2010; Kemp, Perfors, & Tenenbaum, 2007). For example, in a passive learning setting, Lucas and Griffiths (2010) demonstrated that overhypotheses about the functional form transfer to future causal judgments. Participants were asked to identify "blickets" (true causes) from a set of blocks (prospective causes) by observing how different combinations of blocks caused a machine to activate (effect). One group of participants was trained using a *disjunctive* functional form, where the machine activated in the presence of any single blicket; another group was trained using a *conjunctive* form, where the machine activated to two simultaneous blickets. In a subsequent testing phase with new blocks and blickets, both groups of participants observed am-

biguous evidence that was plausible under both disjunctive and conjunctive forms. Crucially, Lucas and Griffiths (2010) found that participants with disjunctive training were more likely to judge a singleton block to be a blicket, while participants with conjunctive training were more likely to judge a pair of blocks to be blickets. In other words, overhypotheses about the functional form constrained participants' causal judgments in a new situation.

Intuitively, it follows that overhypotheses about the functional form can also constrain people's intervention strategies. Returning to our example, if the child has the overhypothesis that multiple batteries are necessary to activate the LED, then they might proceed to test combinations of batteries, rather than individual batteries. Empirically, the past studies described above showed that people can learn overhypotheses about different functional forms and that they can also choose informative interventions based on their causal beliefs. However, to the best of our knowledge, no study has investigated the relationship between the two. Investigating the relationship between causal overhypotheses and active learning is important for improving our understanding of how people's intervention choices can facilitate better causal inferences. Therefore, we explore this relationship through the three questions below.

Question 1: Can people learn overhypotheses about the functional form in an *active* setting? More concretely, if we now allow participants to control their own observations through active interventions, can we still confirm that they learn overhypotheses, as measured by their causal judgments in a new situation? We investigate this question by assigning participants to active training with either the disjunctive or conjunctive form and then measuring their causal judgments about a new task. The new task has either a matched or mismatched form with the training phases. A significant difference in matched vs. mismatched judgments would show that people can successfully learn the overhypothesis from their active training.

This first question serves two purposes. The first is to replicate people's ability to learn different functional forms in *active* settings; this ability was previously established by Lucas and Griffiths's (2010) results in passive settings. In active learning settings, people's interventions tend to be informative, allowing them to learn different causal structures of disjunctive relationships (e.g., Bramley et al., 2015); we expect that their informative interventions will extend to learning about other functional forms, such as conjunctive relationships. Furthermore, Sim and Xu (2017) studied another type of causal overhypothesis—whether objects have matching features (colors and shapes)—and showed that even small children's interventions during free play allowed them to learn overhypotheses. Therefore, by reasoning from these past findings, we anticipate replicating that people *can* learn overhypotheses about the functional form in an active setting.

The second purpose of asking whether people can learn functional forms in an active setting is to consider the negative alternative: Under a time constraint (such as the time limit in our experiment), it is possible that people mainly employ a positive testing strategy (Coenen et al., 2015); when this strategy combines with their prior preference for disjunctive relationships (Lucas & Griffiths, 2010), people might test only singleton objects and, upon observing no positive effects and running out of time, conclude that there are no causal relationships. In this case, they would not learn other overhypotheses like a conjunctive relationship, which is revealed by testing combinations of objects. Therefore, although we anticipate that people will be able to learn overhypotheses in an active setting, we also do not discount the negative alternative.

Question 2: *How* do people choose interventions that help them learn overhypotheses? If we find a positive answer to the first question, we can pursue a more detailed second question about *how* exactly people choose informative interventions to learn about a functional form. We explore this question by measuring the number of objects in each participant's first intervention and their subsequent causal judgment accuracies, predicting that testing more objects leads to better conjunctive judgments and worsened disjunctive judgments.

Question 3: Complementary to Question 2, how do previously learned overhypotheses affect intervention choices in a new situation? Namely, what interventions do people consider informative given a *previously* learned overhypothesis? We explore this question using the same intervention measure as Question 2: the number of objects in each participant's *first* intervention in a new causal learning task, where they need to rely on their previously learned overhypothesis. We predict that a conjunctive overhypothesis will lead to testing more objects than a disjunctive overhypothesis.

To address these questions, we extend Lucas and Griffiths's (2010) blicket experiment to an *active* learning setting. The blicket experiment format allows us to measure causal judgments that indicate whether people learned different functional forms; we additionally measure interventions by allowing participants to interact with objects in the experiment. Finally, we discuss opportunities for improving our experiment design and directions for continuing to investigate the relationship between causal overhypotheses and active learning.

Experiment

Participants 212 participants were recruited using Amazon Mechanical Turk (HIT Approval Rate $\geq$ 99%, Number of HITs Approved $\geq$ 1000, Age $\geq$ 18) for the 8 experimental conditions described in Tables 1 and 2. From top to bottom in Table 2, each condition in the *Disjunctive* column has 27, 29, 29 and 26 participants; each condition in the *Conjunctive* column has 25, 26, 25 and 25 participants. They were paid \$1.5 for completing the study (7.36 minutes on average, excluding the instructions) and received a bonus of up to \$1.05 for their questionnaire performance (mean total compensation: \$2.32).

Design As in Lucas and Griffiths's (2010) experiments, we present participants with blocks and a "blicket machine", and we ask them to identify "blickets" among the blocks by observ-

Table 1: Phase Definitions

Phase	Num. Blocks	Num. Blickets
1 (training)	3	1 (D) or 2 (C)
2 (training)	6	3
3	9	4

Table 2: Experimental Conditions

Training Length	Match with Training Form	Phase 3 Form	
		Disjunctive	Conjunctive
Long	Same	D1 D2 D3	C1 C2 C3
	Different	C1 C2 D3	D1 D2 C3
Short	Same	D1 D3	C1 C3
	Different	C1 D3	D1 C3

Table 1: In each phase, the blickets are a subset of the blocks. In Phase 1, the D (disjunctive) variant has one blicket while the C (conjunctive) variant has two. **Table 2:** Each of the 8 between-participant conditions is a sequence of D/C training phases (1-2) followed by a final D/C Phase 3, representing a manipulation of the *training length*, *match with training form*, and *Phase 3 form*. For example, D1D2D3 is one condition representing a *long* training (D1D2) with the *same* form as the *Phase 3 disjunctive form* (D3).

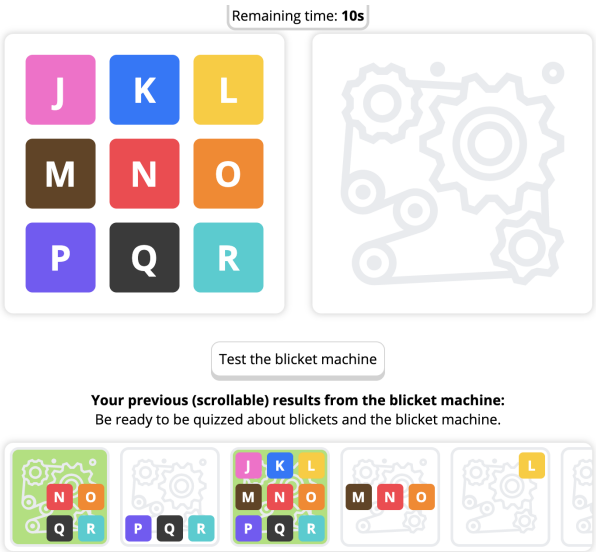

Figure 2: Web interface for our active blicket task (Phase 3). Participants could intervene on the blocks (initially on the left), i.e., click on them to move them on or off the "blicket machine" (right; embellished with cogs) in arbitrary combinations. They could then press a button to test the machine's binary response ("activation" with a green color or "nothing happened" with no color change). Their history of tested combinations was recorded at the bottom, and they could test any number of combinations within a time limit of 45 seconds.

ing the machine's response. Whereas Lucas and Griffiths's study involved fixed and passive sequences of events, ours uses a computer-based presentation that allows participants to produce their own sequences of events through active interventions (see Figure 2; more details in our preregistration[1]). Our experiment contains three phases with successively more challenging tasks, requiring increasingly selective interventions.

To focus on the relationship between overhypotheses and interventions, we only consider simple causal structures and deterministic functional forms. In graphical terms (see Figure 1), all nodes have binary states (0 or 1), blickets are parent nodes, the machine's activation is their common effect, and the non-blicket blocks are isolated nodes. The conditional probability of the machine's activation (i.e., the effect E) is defined by either the **disjunctive form (D)**, where the presence of any blicket (i.e., any parent node $X \in Pa(E)$ with a value of 1) activates the machine $[P(E|Pa(E)) = (\sum_{X \in Pa(E)} X) \geq 1]$, or the **conjunctive form (C)**, where at least two simultaneous blickets activate the machine $[P(E|Pa(E)) = (\sum_{X \in Pa(E)} X) \geq 2]$. Our instructions suggest to participants that the only *structural* problem is to identify the true parent nodes (blickets) that have edges directed toward the blicket machine node. However, to make accurate judgments about the identity of blickets and the behavior of the blicket machine, participants must understand how to make informative interventions given the *functional form* of the causal relationship. For example, blickets can be identified by intervening on single blocks in tasks with a disjunctive form, but this strategy would not reveal any blickets in tasks with a conjunctive form.

Each experimental Phase (Table 1) has a blicket machine with either a disjunctive or conjunctive form. The three phases have the same 45 second time limit for the active task (Figure 2) but different objects and causal structures; later phases increase the structural complexity by adding more blickets (parent nodes) and non-blicket blocks (isolated nodes).

The earlier **Phases 1-2** serve to train participants to successfully learn an overhypothesis about the functional form, e.g., Phase 1 uses a simple three-node causal structure (with 1 blicket for the disjunctive form or 2 for the conjunctive form), where all $2^3 = 8$ possible combinations of blocks can be tested within the time limit. Furthermore, this structure replicates the structure used in Lucas and Griffiths's (2010) experiments, where people have previously succeeded in learning disjunctive and conjunctive overhypotheses from passive data.

The final **Phase 3** serves to investigate how overhypotheses learned from previous training phases inform causal judgments and interventions in this final phase. The task here is more combinatorially complex, involving finding 4 blickets among 9 blocks within 45 seconds. Critically, this complexity increases the importance of relying on overhypotheses from preceding phases, as opposed to finding a brute force solution, e.g., testing all $2^9 = 512$ combinations of blocks, which is impossible under the time limit.

[1]Preregistration: `https://osf.io/n9cx2`

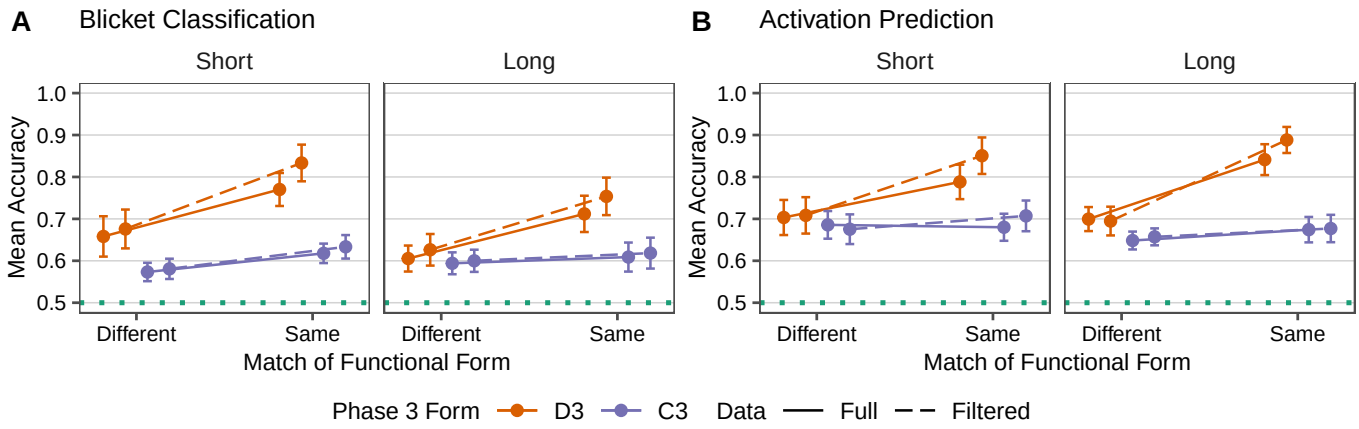

Figure 3: Questionnaire performance in Phase 3 grouped by match, Phase 3 functional form, and training length. The grouping variables are defined in Table 2. Chance (.5) accuracy is shown with a dotted green line. Error bars in either direction denote the magnitude of the standard error. Mean participant accuracies for **A** blicket classification and **B** activation prediction are calculated separately for the full and filtered data.

Procedure Each participant encountered either a *long* training with Phases 1 and 2 or a *short* training with only Phase 1. Following training, they performed the final Phase 3 (see Tables 1 and 2).

Within each phase, participants interacted with the active causal learning task described in Figure 2 and then answered a questionnaire. In the active task, the blicket machine's underlying functional form was either disjunctive or conjunctive. The corresponding questionnaire included two types of causal questions: judgments about whether each block was a blicket or not, and predictions about the activation of the blicket machine in the presence of seven different combinations of blocks. The latter contained combinations with zero, one and two blickets along with other non-blicket blocks, as well as one combination with all blocks in the phase. To allow participants to learn overhypotheses through their own interventions, we only gave between-phase feedback and bonus compensation for activation prediction questions, but not for blicket classification questions.

Overall, we manipulated the number of training phases, the underlying functional form in Phase 3 (disjunctive or conjunctive), and whether the training Phases 1-2 had a matched functional form with Phase 3, creating 8 between-participant experimental conditions (Table 2).

Results

We explored overhypothesis learning in an active setting, focusing on causal judgments and interventions in a new situation—Phase 3. Therefore, we analyzed [2] the questionnaire accuracies and intervention measurements in Phase 3.

Filtering To represent the data accurately while accounting for potential data quality issues, we report results for both the full data set ($N = 212$) and a filtered subset ($N_f = 181$), which includes only participants who made at least 9 interventions in Phase 3. We chose this filtering criterion[3] on the basis that 9 is the minimum number of interventions required to learn the easier variant of the Phase 3 form, i.e. the disjunctive form (Lucas & Griffiths, 2010). Therefore, this filtering aims to consider only participants who were actively engaged with learning the Phase 3 form. From top to bottom in Table 2, each condition in the *Disjunctive* column has 23, 22, 22 and 24 filtered participants; each condition in the *Conjunctive* column has 23, 25, 20 and 22 filtered participants. We address this filtering in the Discussion section.

Active Overhypothesis Training We initially predicted[1] that active training (Phase 1-2) with functional forms would affect participants' causal judgments in Phase 3, facilitating more accurate inferences when the earlier phases have a matched functional form. Conversely, a mismatched form would lead to lower accuracies. A significant difference in the matched and mismatched Phase 3 judgments would show that participants can successfully use active training to learn overhypotheses and apply these overhypotheses in a new situation. We also predicted that this difference would be larger in *long* conditions, where the matched or mismatched overhypothesis is reinforced with more training.

In other words, we expected the match of overhypotheses from preceding phases (same or different compared to Phase 3's functional form), the length of training (*long* or *short*), and their interaction to be predictors of Phase 3 questionnaire performance, considering the Phase 3 functional form (disjunctive or conjunctive) as a covariate. Thus, we used these variables to fit two logistic regression models to

[2] Analysis code: `https://github.com/chen10an/active-blicket-comp/tree/1.0.x/analysis`

[3] In an earlier filtering approach, we considered only 116 participants who judged at least one block as a blicket and made at least one intervention in Phase 3. This yielded similar results to our current inferential statistics. We use the current approach to exclude fewer participants and to avoid conditioning our filtering on participant judgments, as suggested by anonymous reviewers.

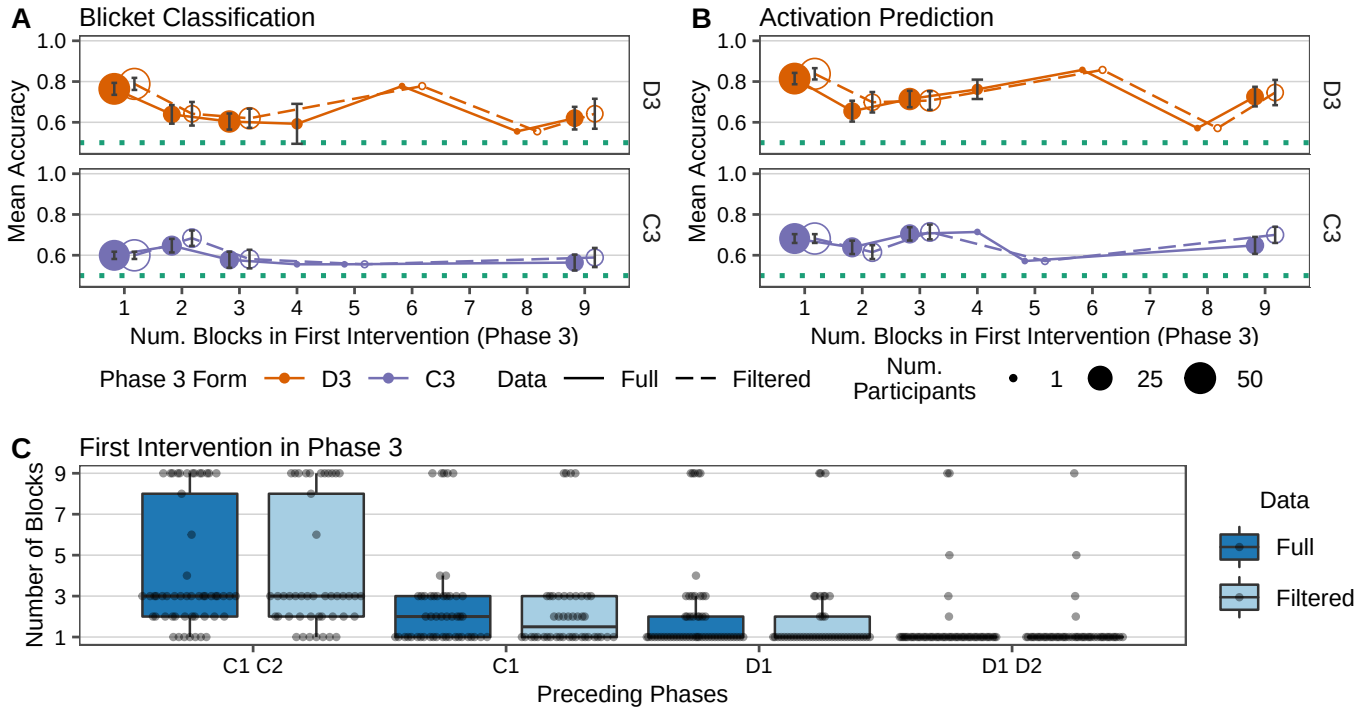

Figure 4: Judgments and interventions in Phase 3. Mean participant accuracies for **A** blicket classification and **B** activation prediction questions in Phase 3 are grouped by the number of blocks in the first intervention and the Phase 3 functional form. The mean is calculated separately for the full (solid lines) and filtered (dashed lines) data. Error bars in either direction denote the magnitude of the standard error but are omitted for points with a single participant, where the standard error is ill-defined. Chance (.5) accuracy is shown with a dotted green line. **C** Number of blocks tested in the first intervention in Phase 3, compared to the functional form of preceding phases. The box-and-whisker plots show the quartiles of the full or filtered data; each overlaid point is a participant.

predict the per-participant accuracy percentage in Phase 3 blicket identification (binomial with 9 trials) and activation prediction (binomial with 7 trials) questions, respectively. For blicket classification accuracy, we confirmed a significant main effect of the match of overhypotheses ($z = 2.62, p = .009$; filtered: $z_f = 3.25, p_f = .001$). For activation prediction accuracy, the main effect of *match* was not significant in the full data ($z = 1.24, p = .215$), but was significant for the filtered participants who were more actively engaged with the Phase 3 task ($z_f = 2.53, p_f = .012$). Surprisingly, the length of training and its interaction with the match of preceding phases were not significant predictors for either type of question (all $p \geq .326$). Consistent with past results suggesting that people find disjunctive relationships easier to learn (Lucas & Griffiths, 2010), the Phase 3 functional form had a significant main effect for both blicket classification ($z = 3.99, p < .001$; filtered: $z_f = 4.78, p_f < .001$) and activation prediction ($z = 3.67, p < .001$; filtered: $z_f = 4.23, p_f < .001$).

The non-significant effect of the training length interaction in both models may be attributable to weak *match* effects in Conjunctive Phase 3 (C3) conditions. Therefore, we used Welch t-tests (two-tailed) to investigate the specific effects of *match* between pairs of conditions, where each comparison is also visualized in Figure 3. The Disjunctive Phase 3 (D3) comparisons were mostly consistent with our predictions that matching overhypotheses would improve performance: in the full data, the mean blicket accuracy showed a trend toward improvement from mismatched to matched conditions with *long*, $t(47.82) = -2.00, p = .051$, and *short* training, $t(49.76) = -1.80, p = .078$. These trends became significant improvements in the filtered data (long: $t_f(42.10) = -2.18, p_f = .035$; short: $t_f(43.99) = -2.47, p_f = .017$). The D3 activation prediction accuracy also improved significantly from mismatched to matched conditions with *long* training, $t(50.00) = -3.04, p = .004$ (filtered: $t_f(42.37) = -4.18, p_f < .001$). The *short* training improvement was not significant in the full data ($t(52.70) = -1.45, p = .154$), but was significant in the filtered data ($t_f(43.89) = -2.31, p_f = .025$). In the *long* and *short* C3 conditions, however, the difference between matched and mismatched accuracies was non-significant for both types of questions (all $p \geq .164$). It is possible that even with a matching overhypothesis and longer training, the C3 task was too difficult, which could account for the non-significant effect of the training length interaction in our modeling. We address C3's difficulty in the Discussion section.

How Interventions affect Overhypotheses To explore how intervention choices affect overhypothesis learning, we measured the number of blocks in the *first* intervention of

Phase 3 and analyzed how this intervention affected subsequent causal judgment accuracies. We predicted that choosing more objects would lead to higher accuracies in Conjunctive Phase 3 (C3) and decreased accuracies in Disjunctive Phase 3 (D3). Figure 4A shows this effect in the blicket judgment accuracy, especially in the filtered data. Specifically, the disjunctive blicket accuracy has a decreasing trend as the number of blocks increases, while the conjunctive blicket accuracy peaks at two blocks. Figure 4B suggests that the relationship between the number of blocks and activation prediction performance is more complex.

To investigate the trends in Figure 4A-B, we fitted two (binomial) logistic regression models to predict blicket classification (9 trials) and activation prediction (7 trials) accuracy, respectively. The predictors included the number of blocks tested in the first Phase 3 intervention, the functional form of the Phase 3 relationship (disjunctive or conjunctive), and their interaction. The model results were largely consistent with our previous observations in Figure 4A-B: For blicket classification accuracy, there was a significant main effect of the Phase 3 functional form ($z = 4.26, p < .001$; filtered $z_f = 4.90, p_f < .001$), underscoring the relative difficulty of the conjunctive condition, and no significant main effect for the number of blocks (all $p \geq .382$), suggesting that any effect of the number of blocks was not due to choosing more (or fewer) blocks being a better general-purpose policy. Rather, the effect of the number of blocks was due to being informative of a particular Phase 3 form: this interaction did not reach significance in the full data ($z = -1.73, p = .084$), but was significant for the more engaged participants in the filtered data ($z_f = -2.02, p_f = .043$). For activation prediction accuracy, there was also a significant main effect of the Phase 3 functional form ($z = 3.38, p < .001$; filtered: $z_f = 4.29, p_f < .001$), but no other significant effects (all $p \geq .078$). Figure 4A-B suggests that, even though participants were able to identify a larger subset of blickets with efficient interventions (A), this partial knowledge was not sufficient to perform better in the activation prediction questions (B), which had a larger coverage over blickets and their combinations with other blocks.

How Overhypotheses affect Interventions We also used the first Phase 3 intervention to explore how previously learned overhypotheses about the functional form shaped interventions in a new task. The first Phase 3 intervention occurred before participants learned anything about the functional form in Phase 3, making it a simple marker of how their interventions were informative under an overhypothesis learned from *previous* phases. Under a disjunctive overhypothesis, testing individual blocks would be a straightforward and efficient way to identify blickets, requiring only nine interventions in all. In contrast, testing individual blocks would be completely uninformative under a conjunctive overhypothesis. This intuition is consistent with the trends in Figure 4C.

To further test the trends in Figure 4C, we used a linear model to predict the number of blocks in the first intervention, where the predictors were the functional form in the previous training phases (disjunctive or conjunctive), the length of training (*long* or *short*), and their interaction. There was a significant interaction effect ($t(205) = -3.59, p < .001$; filtered: $t_f(177) = -3.46, p_f < .001$) and significant main effect of training length ($t(205) = 3.38, p < .001$; filtered: $t_f(177) = 3.66, p_f < .001$). The non-significant main effect of the preceding functional form ($p \geq .322$ for both the full and filtered data) may be attributable to weaker effects in the short conditions—see the C1 and D1 box-and-whisker plots in Figure 4C. Consistent with these model results and with the trends in Figure 4C, the mean number of blocks tested in the first intervention was significantly higher after conjunctive training (C1 and C1C2) than after disjunctive training (D1 and D1D2), $t(183.39) = 4.62, p < .001$ (filtered: $t_f(144.96) = 4.75, p_f < .001$).

Similarly, in the free-text responses, participants often reported intervention strategies that were shaped by the functional form they had learned in previous phases: participants with longer disjunctive training tended to test "each block individually knowing that it would activate whether alone or with others", while those with longer conjunctive training tended to test more blocks at once, e.g., "groups of three to find blicket pairs, groups of two to narrow it further, known blickets against the remaining unknown single blocks".

Discussion

Our experiment explored three questions about the relationship between overhypotheses about the functional form and people's interventions during active causal learning. First, we asked whether people can learn overhypotheses about the functional form in an active setting: we found positive results in the final disjunctive phase, where people's judgment accuracies largely showed they had succeeded in learning the functional form from their active training. Their judgments were more accurate following a matched training (disjunctive) but less so following a mismatched one (conjunctive). Second, we asked *how* people's interventions help them learn about the functional form: we measured the number of blocks in the first intervention of the final phase, finding that this intervention measure predicted blicket judgments in the filtered data, where participants were more engaged with the task. While a singleton block indicated better blicket judgments about disjunctive relationships, a pair of blocks indicated better blicket judgments about conjunctive ones (Figure 4A). Finally, we asked the complementary question about how people's previously learned overhypotheses shape their intervention choices in a new task: using the same intervention measure in the last phase, we found that participants who trained with a disjunctive functional form in previous phases predominantly started by testing a singleton block, while those who trained with a conjunctive form tested more blocks; this pattern became more apparent with longer training (Figure 4C). Overall, these results begin to expand our understanding of causal interventions to a new perspective—overhypotheses about the functional form of causal relationships.

In addressing whether active training helps people learn about the functional form, there were also surprising results. We expected a strong effect of active training (matched vs. mismatched) in Conjunctive Phase 3 (C3) judgments, especially in the *long* C3 conditions that gave additional training opportunities to learn about the underlying functional form. Instead, we found a non-significant trend. One possible explanation would be that participants were not learning the conjunctive overhypothesis through active training (i.e., C1 or C1C2), and thus, participants with a matched conjunctive training were performing no better than those with a mismatched disjunctive training. However, this explanation seems unlikely in light of a small additional study we conducted. In that study, we used the same dependent measure as Lucas and Griffiths (2010) to test whether participants learned a conjunctive overhypothesis (0-10 ratings of object D being a blicket; see Lucas and Griffiths's paper for details) after training in our active C1. We compared this to our active D1 training as well as our replication of the passive disjunctive training in Lucas and Griffiths's Experiment 2. There was a significant effect of C1 ($M = 7.40$, $SD = 2.30$) vs. disjunctive ($M = 1.25$, $SD = 2.50$) training, despite a small sample size, $t(6.28) = 3.80$, $p = .008$ (two-tailed Welch). This result suggests people have little trouble learning conjunctive overhypotheses during C1 training. Therefore, the non-significant effect of *match* in C3 conditions was likely due to other factors.

These factors might include participant fatigue or frustration during C3, regardless of matched or mismatched training. For example, participants were not given enough time (45s) to test all 36 pairs of blocks. In contrast, the same time limit allowed participants to gather exhaustive information about D3 by testing all 9 singleton blocks. Indeed, performance was lower in C3 conditions than in D3 conditions ($p < .001$ for both blicket classification and activation prediction questions in the full and filtered data). The lower C3 performance might be further compounded by the factor that motivated our data filtering ($N_f = 181$ remained in the filtered data, out of $N_{\text{total}} = 212$): lack of participant engagement for making at least 9 interventions during the active learning task. These factors are limitations that follow-up studies may address by using different schemes for resource limits (e.g., requiring a constant number of total interventions and/or lowering the number of blocks) and improving the in-experiment engagement checks.

Finally, future work is needed to expand this paper's answers to the questions about *how* interventions and functional forms influence each other. For example, can we characterize people's intervention strategies beyond counting the number of blocks? Are these intervention strategies consistent with the goal of maximizing information gain about the functional form? Addressing these future questions will require extending the kinds of Bayesian models developed in previous active causal learning studies (Bramley et al., 2015; Steyvers et al., 2003; Coenen et al., 2015).

Acknowledgments

We thank anonymous reviewers for their helpful suggestions[3].

References

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.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349.

Gopnik, A., & Sobel, D. M. (2000). Detecting Blickets: How Young Children Use Information about Novel Causal Powers in Categorization and Induction. *Child Development*, *71*(5), 1205–1222.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, *35*(8), 1407–1455.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2016). A Bayesian Theory of Sequential Causal Learning and Abstract Transfer. *Cognitive Science*, *40*(2), 404–439.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological review*, *115*(4), 955.

Lucas, C. G., Gopnik, A., & Griffiths, T. L. (2010). Developmental differences in learning the forms of causal relationships. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32).

Lucas, C. G., & Griffiths, T. L. (2010). Learning the Form of Causal Relationships Using Hierarchical Bayesian Models. *Cognitive Science*, *34*(1), 113–147.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

Sim, Z. L., & Xu, F. (2017). Learning higher-order generalizations through free play: Evidence from 2- and 3-year-old children. *Developmental Psychology*, *53*(4), 642–651.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.