

Supplementary material for paper "Building Object-based Causal Programs for Human-like Generalization"

A PCFG for symbolic causal functions

Let f be a causal function that takes the agent object a and the recipient object r as input, and outputs the result object r' . We assume a prior over possible causal functions in the form of a PCFG $\mathcal{G} = (\Gamma, \Theta)$, where Γ is a set of production rules and Θ is a set of production probabilities (see Table 2). Let ϕ_i denote the i -th feature in the set of all observable object features Φ . Grammar \mathcal{G} thus generates expressions that specify features of the result object. For example, if one of the features is "color", a possible causal function could be $\text{color}(r') \Leftarrow \text{red}$ — recipient will turn red — or $\text{color}(r') \Leftarrow \text{color}(a)$ — recipient will take the agent's color, and so on. The grammar is set up to allow for arbitrarily complex expressions through the "bind additional" production rule (Table 2, row 2), allowing a rule to produce conjunctions of feature changes, for example, $\text{AND}(\text{color}(r') \Leftarrow \text{red}, \text{shape}(r') \Leftarrow \text{triangle})$.

We assume that any features unspecified by a causal function follow the principle of inertia, and remain as they were before the causal interaction.

Note that although grammar \mathcal{G} is very similar to a PCFG, it is not context-free strictly speaking: the "bind feature" production rule (Table 2, row 1) binds a feature to a lambda expression, and the subsequent steps within the scope of the λ -abstraction all refer to this feature.

For simplicity, we assume uniform transition probabilities for each production rule. i.e., $\theta_l = \frac{1}{|I|}$ for each row I with production rules $l \in I$. By design, this grammar is inherently more likely to produce simpler expressions. The "bind additional" rule is called with probability 0.5, and thus the number of conjunctions in the final expression follows a geometric decay with only 50% combining more than one assertion, 25% containing more than two, and so on. The prior for a given expression is thus simply the product of all the productions that produced it:

$$P_{\mathcal{G}}(f) = \prod_{l \in \Gamma} (\theta_l)^{c_l} \quad (6)$$

where $l \in I$ is the transition probability for production rule $l \in \Gamma$, and c_l is how many times rule l was used for generating this causal function.

A causal function outputs result object(s) when particular agent and recipient objects are provided. Take $\text{AND}(\text{color}(r') \Leftarrow \text{color}(a), \text{shape}(r') \Leftarrow \text{square})$ for example. For an agent a that is a red-circle and a recipient r that is a blue-pentagon, r will become r' : a red-square. When a causal function f involves a negation, it could have produced more than one outcome. For example let w be $\text{shape}(r') \Leftarrow \neg \text{triangle}$, any object that is not triangular (and share the same color as r) is a possible option for being r' . We further assume for simplicity that the different potential outcomes are equally probable, and thus likelihood of a data point $d = (a, r, r')$ generated by a causal function f is given by

$$P(d|f) = P(r'|f, a, r) = \begin{cases} \frac{1}{D(f(a, r))} & \text{if } r' \in D(f(a, r)), \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where D stands for *domain* and $D(f(a, r))$ refers to the set of all possible result objects coming out of f given agent a and recipient r . We initially assume a likelihood to 0 for any observation a, r, r' incompatible of $f(a, r)$, but later consider "soft" variants in which functional relationships are somewhat fallible.

This framework naturally favors deterministic causal functions that are consistent with the evidence: if a causal function predicts a specific result, when that outcome is indeed observed, likelihood will be 1. In contrast, a causal function that predicts a range of outcomes will inevitably assign a lower likelihood to any one of these.

Table 2: Example probabilistic grammar \mathcal{G}

| Production rules | | | |
|--------------------|-----------------|------------------------------|--------------------------------|
| Bind feature | $S \rightarrow$ | $\lambda_{\phi_i} : A, \Phi$ | |
| Bind additional | $A \rightarrow$ | B | $\text{AND}(B, S)$ |
| Relation | $B \rightarrow$ | $\phi_i(r') \Leftarrow C$ | $\phi_i(r') \Leftarrow \neg C$ |
| Reference | $C \rightarrow$ | D | E |
| Relative reference | $D \rightarrow$ | $\phi_i(a)$ | $\phi_i(r)$ |
| Absolute reference | $E \rightarrow$ | value^{ϕ_i} | |

Note: ‘‘Bind feature’’ samples a feature without replacement from the set of all features. ϕ_i in D uses the feature selected in A , and value in E is sampled uniformly from the support of the feature selected in A .

B Unpacking latent causal categories

A CRP is a stochastic process widely used for creating partitions among entities [47]. It draws on an analogy of sequentially seating infinite incoming customers to infinitely many tables in a Chinese restaurant, where each table is also of infinite capacity. The first observation $d^{(1)}$ is always assigned the first category $z^{(1)}$; when $i > 1$, the probability for assigning category $z^{(i)}$ is given by

$$P(z^{(i)} = x | z^{(-i)}) = \begin{cases} \frac{\alpha}{i-1+\alpha} & \text{if } x \text{ is a new category} \\ \frac{|z^{(j)}|}{i-1+\alpha} & \text{if } x = z^{(j)} \end{cases} \quad (8)$$

where $z^{(j)}$ is an existing category, and $|z^{(j)}|$ is the number of assigned objects in category $z^{(j)}$. Parameter α ($\alpha > 0$) is known as the concentration, or dispersion parameter—the larger α is, the more likely a new object falls into a new category. Holding the same α , categories with more members are preferred as they seem to be more ‘‘common’’.

Objects in a category characterize shared feature similarities, modeled by a multinomial distribution over a finite number of feature values. Let $\mu^{(z_i)} = [\mu_1, \dots, \mu_n]$ be the mean feature vector of a given category z_i , where each subscript k in μ_k refers to a feature, μ_k is the mean value of feature k for all the objects in category z_i , the probability that an object is assigned to a particular category according to feature similarities is given by

$$P(o^{(i)} | \mu^{(z_i)}) = \prod_{k=1}^n \text{Bernoulli}(o^{(i)}; \mu_k) \quad (9)$$

To compute μ_k , let $o_v = [o_{v_1}, \dots, o_{v_n}]$ be the feature values of an object o , where each v represents a feature value, $o_{v_i} = 1$ if object o has this feature value and $o_{v_i} = 0$ otherwise. For a category $z = \{o^{(i)}, \dots, o^{(m)}\}$, $z_v := \sum_{j=1}^m o_{v_i}^{(j)}$, which can be written as $z_v = [z_{v_1}, \dots, z_{v_n}]$, where $z_{v_i} = \sum_{j=1}^m o_{v_i}^{(j)}$. Mean feature $\mu_k := \frac{z_{v_k}}{\sum_{l=1}^n z_{v_l}}$. We assign a Dirichlet prior to this multinomial distribution in order to capture how important feature similarity is in forming categories. Without leaning towards any specific feature, the prior distribution over mean features is simply $\text{Dir}(\beta)$, $\beta \geq 0$.

It is not obvious whether mean features should be drawn from the agent object, recipient object, or both, therefore we introduce one more hyper parameter γ , referring to the probability that mean feature is purely based on the agent: when $\gamma = 1$, categorization is only grounded on the agent objects, when $\gamma = 0$, only recipient’s features are considered for categorization, and when $\gamma = 0.5$, both agent and recipient are considered equally. We thus consider $0 \leq \gamma \leq 1$.

In sum, a Dirichlet Process that creates a distribution over causal category distributions according to Equation 2 has the priors:

$$\begin{aligned} z^{(i)} | z^{(-i)} &\sim \text{CRP}(\cdot | \alpha) \\ \mu^{(i)} &\sim \text{Dir}(\cdot | \beta) \\ f^{(z_i)} &\sim \mathcal{G}(\cdot) \end{aligned} \quad (10)$$

Algorithm 1 Process model

```
1: Initialize an empty list of causal categories  $Z$  ▷ Initialization
2: Assign  $a^{(0)}, r^{(0)} \in d^{(0)}$  to category  $z^{(1)}$ , update  $\mu^{(1)}$  ▷ Learning example goes to the first category
3: Sample  $f^{(1)}$  from the learning posterior
4: Record  $z^{(1)}$  in list of causal categories  $Z$ 
5: for each  $d^{(i)} \in D_G$  do
6:   sample  $z^{(i)} \propto P(z^{(i)}|z^{(-i)})P(a^{(i)}, r^{(i)}|\mu^{(z_i)})$  ▷ Equation 5
7:   if  $z^{(i)} \in Z$  then ▷ If current object belongs to an existing category
8:      $r'^{(i)} \sim f^{(i)}(a^{(i)}, r^{(i)})$  ▷ Make prediction
9:     Add  $a^{(i)}, r^{(i)}$  to  $z^{(i)}$ : update  $\mu^{(i)}$  ▷ Update  $Z$ 
10:  else
11:    Assign  $a^{(i)}, r^{(i)}$  to a new category  $z^{(k)}$ : update  $\mu^{(k)}$  ▷ Create a new category
12:    Sample  $f^{(k)}$  from the prior
13:     $r'^{(i)} \sim f^{(k)}(a^{(i)}, r^{(i)})$  ▷ Make prediction
14:    Add  $z^{(k)}$  to  $Z$  ▷ Update  $Z$ 
15:  end if
16: end for each
```

And likelihoods are given by

$$\begin{aligned} a^{(i)}, r^{(i)} | \mu^{(z_i)} &\sim \text{Dir}(\cdot | \mu^{(z_i)}, \beta) \\ d^{(i)} | f^{(z_i)} &\sim f^{(z_i)}(a^{(i)}, r^{(i)}) \end{aligned} \tag{11}$$

where $\mu^{(z)}$ is the mean feature vector, and $f^{(z)}$ the assigned causal function.

To approximate the posterior with Gibbs sampling, we construct a chain of samples where for each iteration, we sample a causal category for a random observation $d^{(i)}$ while fixing the category assignment to the other observations, and a sampled causal category $z^{(i)}$ will then update the category parameters $\mu^{(z_i)}$ and $f^{(z_i)}$. The category sampling step of this Gibbs sampler follows Equation 2, and the local parameter update step follows definition of computing these parameters given objects in this category. When the number of iterations $n \rightarrow \infty$, the sampled categories \tilde{Z}_n converges to the true posterior.

C Process variant

The process model first assigns the object-pair in the learning example to an initial causal category $z^{(1)}$ governed by a causal law sampled from the posterior distribution $P(f|d)$. Crucially, for each generalization task, it then assigns the encountered object pair scenario to either an existing causal category or a new category according to Equation 5. If an existing causal category is selected, the model simply applies the requisite causal law category to make its prediction. If a new category is sampled, however, a new causal law will be assigned to this category. Since there is no evidence about what causal law may apply to this new category, this new causal law is sampled from the prior. Algorithm 1 shows this process.

Instead of approximating a posterior over infinitely many possible categories as the normative model, the process model maintains a small set of available categories that are created online as new generalizations are performed. Furthermore, after categorizing an observation, the process model updates the list of causal categories Z with this categorization decision, reflecting a commitment to its earlier decisions. Hyperparameter α thus plays a slightly different role in the process model. When $\alpha \rightarrow 0$, the process model becomes increasingly likely to stick with existing categories (Equation 8).

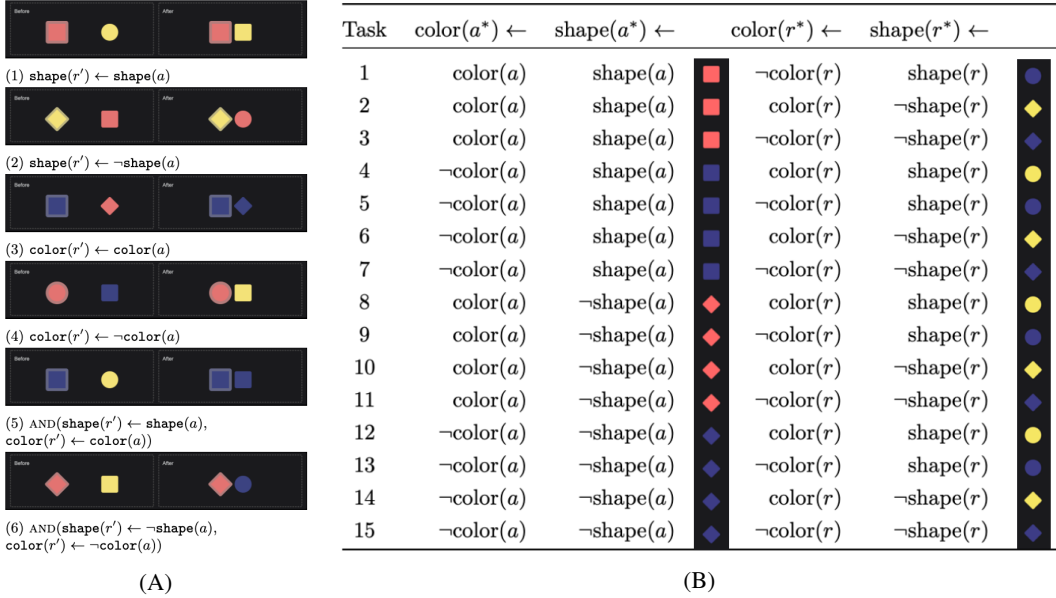


Figure 5: Experiment 1 stimuli. (A): Learning conditions, showing objects before and after a causal interaction. (B): Generalization task configurations a^* , r^* are the agent and recipient in each generalization task; a and r are the agents and recipients in the learning example. Example stones are for learning condition A1.

D Experiment 1 setup

D.1 Stimuli and design

Participants were told that they were making predictions about the behavior of a magic world containing magic stones (agents) and normal stones (recipients). In short videos, participants observed a magic stone collide with a normal stone and appear to alter the normal stone’s color and/or shape (see Figure 1). Magic stones had a thick border while normal stones had no border. We manipulated two object features—color $\{\text{red}, \text{yellow}, \text{blue}\}$ and shape $\{\text{circle}, \text{square}, \text{diamond}\}$, leading to $3 \times 3 = 9$ possible configurations for each object and a nominal $9 \times 9 \times 9 = 729$ configurations of agent and of recipient both pre- and post- the causal interaction.

We used a 6×2 between-subject design. There were six learning examples varied between subjects (Figure 5A)—each participant saw one. Each learning example demonstrates a causal effect differing in whether it results in a change to one or both features of the recipient object, and whether either or both of these new values match the agent object’s features. Note that the function descriptions were not shown to participants and are by no means the only possible way to characterise the causal relationship being displayed.

For each learning example, we constructed 15 generalization tasks by varying object features systematically from the learning example (Figure 5B). For example, A1 (see Figure 5A) depicts a *red square* agent and a *yellow circle* recipient, and according to the specifications in Figure 5A, task 1 for A1 has a *red square* agent, and a *blue circle* recipient. We call the sequence of tasks from 1 to 15 “near-first transfer” because this sequence of tasks starts with those that differ by only one feature from the learning example and progress to scenes in which all of the features differ. Conversely, we call the sequence of tasks 15 to 1 the “far-first transfer” sequence, because it starts with sets of stones that are completely different from those in the learning examples and progresses back to the more similar cases. Within each sequence, whether the set of different-color tasks or the set of different-shape tasks appeared first (task 1 & 2, 5 & 6, 9 & 10, 13 & 14, 4—7 & 8—11) was shuffled to counterbalance feature order.

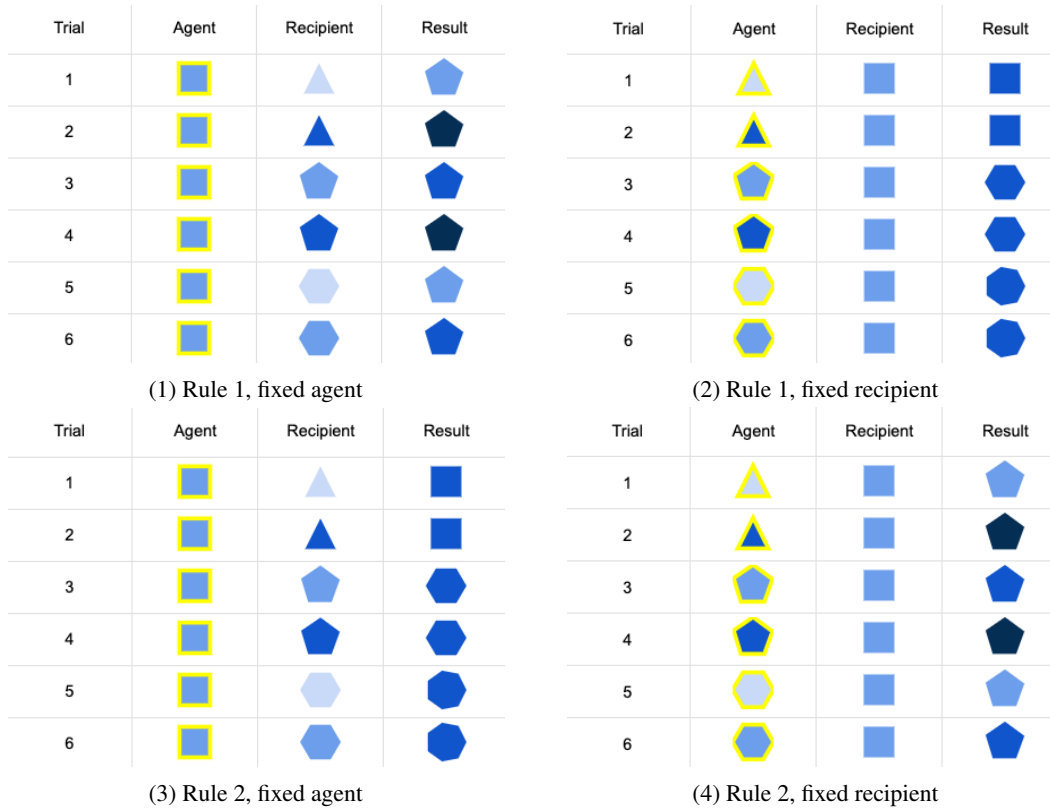


Figure 6: Experiment 2 learning conditions.









D.2 Procedure

After instructions, participants had to pass a comprehension quiz to start the main task. The main task contained a learning phase and a generalization phase. During learning, participants watched one specific magic stone’s effect on a normal stone (Figure 1A–C), and they could replay the effect as many times as they wanted. After that, participants were asked to make predictions for 15 new pairs of magic stones and normal stones sequentially, by selecting from a panel of 9 possible stones (Figure 1D). A summary of the learning example (as used in Figure 5A) was displayed at all times and the animation was replayed once between each generalization task to ensure it was not forgotten. A demo of the task is available at http://bramleylab.ppls.ed.ac.uk/experiments/bnz/magic_stones/index.html.

D.3 Model fits

Both the UnCala and LoCaLa models were fit to the behavioral data using the `optim` function in R. As for the LoCaLaPro model, since it approximates posterior distribution with simulation-based method, we optimized parameter values via grid search. Firstly, we set up a coarse grid with $\alpha = \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.5, 2, 4, 8\}$, $\beta = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$. After running this coarse grid and locating an optimal area, we ran another search over a finer grid for $\alpha = \{0.28, 0.30, 0.32, 0.34, 0.36, 0.38, 0.4, 0.42, 0.44, 0.46, 0.48, 0.5, 0.52\}$ (β is the same as previously) to improve precision.

Table 3: Experiment 2 generalization task configurations

| | For the fixed object | Instance | For the varied object | Instance |
|---------|---|---|----------------------------------|---|
| $o^* =$ | shade(o), edge(o) |  | shade(o), \neg edge(o) |  |
| | \neg shade(o), edge(o) |  | |  |
| | shade(o), \neg edge(o) |  | \neg shade(o), edge(o) |  |
| | \neg shade(o), \neg edge(o) |  | |  |

o^* is the object in generalization tasks, o is the object shown during learning. For the varied object, \neg shade(o) means picking a shade that has not appeared during the learning phase, and we chose two instances for it.

E Experiment 2 setup

E.1 Stimuli and design

Similar to Experiment 1, we varied the shape and color properties of the objects. However, instead of using categorical values, we introduced intuitively ordinal feature values. Shapes were all equilateral and differed in terms of their number of sides: 3 (triangle), 4 (square), 5 (pentagon), 6 (hexagon), and 7 (heptagon); colors were of identical hue and saturation (blue) but differed in lightness varying between: 1 (light blue #c9daf8), 2 (medium blue #6d9eeb), 3 (dark blue #1155cc), and 4 (very dark blue #052e54). Staying within the features’ observed values this leads to $4 \times 4 = 16$ possible configurations for each object, and a nominal $16^3 = 4096$ possible configurations for objects both pre- and post- the causal interaction. These ordinal features enlarge the space of effects and greatly enriches the space of plausible rules, for example allowing causal laws in which a recipient stone becomes *darker* or *lighter* when acted upon, gaining or losing sides, as well as those involving copying or taking specific or random values.

During learning, each participant observed six causal interactions between different pairs of agent and recipient before making generalizations. We included 2 (evidence-balance) \times 2 (ground truth) between-subject factors (see Figure 6). with respect to evidence-balance, for *fixed-agent* conditions B1 and B3, an identical agent was shown in all learning examples, while the recipients it acted on were varied systematically; in the *fixed-recipient* conditions B2 and B4, the recipient object was always identical but was acted on by six different agents. We designed the evidence to be consistent with two “ground truth” rules that counterbalance between the roles of the shape and the color features:

Rule 1 (B1/B2) The recipient gets one increment darker and takes the agent’s shape plus one edge
 $\text{AND}(\text{edge}(r') \leftarrow \text{edge}(a) + 1, \text{shade}(r') \leftarrow \text{shade}(r) + 1)$

Rule 2 (B3/B4) The recipient gains an edge and takes the agent’s shade plus one shade increment
 $\text{AND}(\text{shade}(r') \leftarrow \text{shade}(a) + 1, \text{edge}(r') \leftarrow \text{edge}(r) + 1)$

Note that these “ground truth” rules are just one of an unbounded set of possible universal causal relations consistent with the six learning trials, and a single universal category is just one of a much larger set again of possible local causal law category structures.

We composed generalization tasks according to the configurations in Table 3. In total there were $4 \times 4 = 16$ generalization tasks for each condition.

E.2 Procedure

After completing instructions, participants had to pass a comprehension quiz to proceed to the main task, consisting of a learning phase, self-report, and a generalization phase. After the main task, participants provided demographic information and feedback. A demo of the task is available at <http://bramleylab.ppls.ed.ac.uk/experiments/bnz/myst/p/welcome.html>.

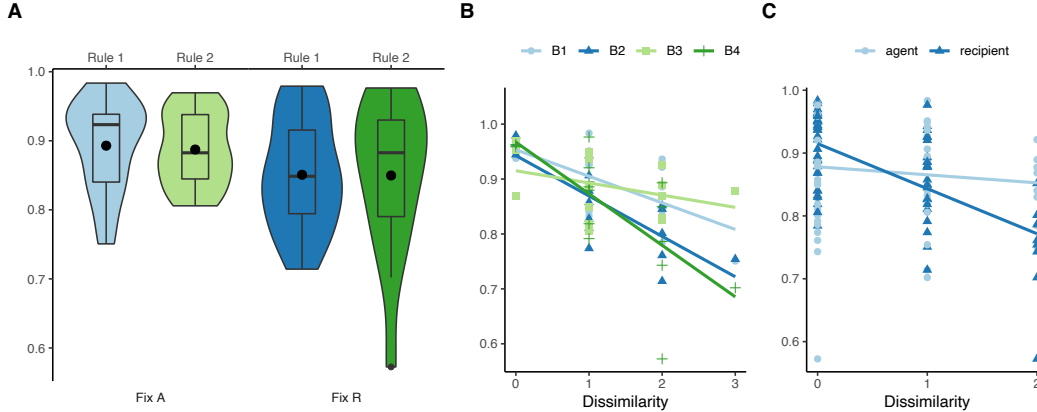


Figure 7: Behavioral results of Experiment 2. All y -axes are Cronbach’s alpha values. A. Task-wise inter-person consistency per condition. Violin plots are density. Black dots are mean Cronbach’s alpha values per condition. The major bar in the box plot is median and box extent is the 25 and 75 quantiles. B. Inter-person consistency per task differences. C. Inter-person consistency per role differences.

Each participant was randomly assigned to one of the four learning conditions (Figure 6). The six pairs of agent and recipient stones were shown in random order, one after another. By clicking a “Test” button, they could watch the causal interaction as many times as they wanted. After each object pair was tested, a summary visualization of the agent, recipient and the result was added to the top of the page (see Figure 1E–F), and remained visible for the rest of the task. After the learning phase, participants were asked to write down their best guess about how the mysterious stones worked, and told they would receive a \$0.50 bonus if they described the true underlying causal law. In the generalization phase, participants faced the 18 generalization trials sequentially in random order. For each, participants predicted the result recipient by selecting a number of edges and the shade of blue from two drop-down menus (see Figure 1F). Participants were instructed they would receive a \$0.10 for each correct prediction. We bonused participants as described afterwards.

E.3 Generalization consistency

As with Experiment 1, we measured inter-person consistency in generalization predictions computing ρ_T for the sixteen generalization tasks per condition (excluding the two catch-trials), totalling $4 \times 16 = 64$ values. Mean consistency was $\rho_T = 0.87 \pm 0.08$, with min $\rho_T = 0.57$, max $\rho_T = 0.98$. To compare generalization consistency against random selections, for each condition we conducted Fisher’s exact test on the contingency table of selecting each possible result per trial. For all four conditions, $p < 0.001$. Thus, as in Experiment 1, participants produced systematic generalization patterns.

We then compared inter-person generalization consistency by condition. As illustrated in Figure 7A, the *fixed-agent* condition induced higher consistency ($\rho_T = 0.89 \pm 0.06$) than the *fixed-recipient* condition ($\rho_T = 0.85 \pm 0.1$), $t(31) = 2.12$, $p = 0.04$, 95%CI = [0.001, 0.08], while the difference in ρ_T between the ground truth condition was negligible, $t(31) = 0.22$, $p = \text{n.s.}$. No interaction was detected. In short, participants made more homogeneous predictions after observing the same agent acting on a range of recipients, and diverged more having observed different agents interacting on the same recipient.

Generalization consistency decreased as objects in the generalization tasks become more distinct from those in the learning examples (Figure 7B). To show this, we constructed a rough measure of *dissimilarity*, by counting the features of generalization trials that took novel values never observed in the learning phase. Formally, let F_L be the set of unique feature values of all the objects appeared during learning, and F_i be the set of unique feature values of objects in a generalization trial i , dissimilarity score $DS = |F_i \setminus F_L|$. By design, dissimilarity scores $DS \in \{0, 1, 2, 3\}$ (Table 3). We found a significant negative relationship between task dissimilarity and generalization consistency, $\beta = -0.06$, $F(1, 62) = 37.48$, $p < 0.001$.

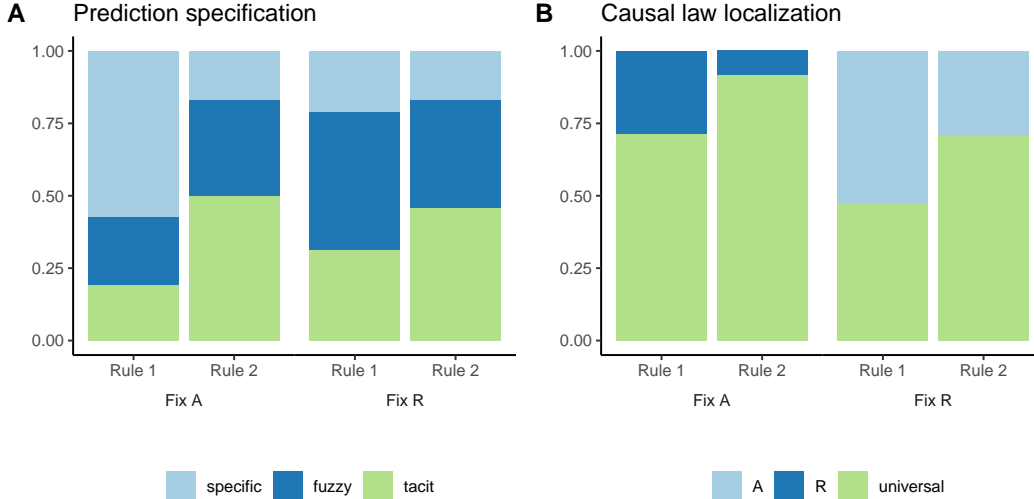


Figure 8: Rule guess categories.

Finally, we fit a linear regression model predicting ρ_T with task dissimilarity, evidence-balance and ground truth, $F(3, 60) = 15.63, p < 0.001$. This revealed main effects of dissimilarity ($\beta = -0.06, p < 0.001$) and evidence-balance (*fixed-recipient*, $\beta = -0.04, p = 0.01$), but not ground truth (*rule 2*, $\beta = -0.003, p = \text{n.s.}$). As depicted in Figure 7B, consistency of judgments in the fixed-agent conditions (B1 & B3, lighter lines) decreased slower than the fixed-recipient conditions as dissimilarity increased (B2 & B4, darker lines).

Not only did the evidence-balance condition have a significant effect on generalization consistency, dissimilarity of the agent or recipient objects in the generalization tasks was also associated with lower consistency (Figure 7C). Holding recipient dissimilarity constant, increasing agent dissimilarity does not predict prediction consistency significantly, $F(1, 62) = 0.77, p = \text{n.s.}$; however, recipient dissimilarity does, $F(1, 62) = 38.8, p < 0.001$.

E.4 Self-reports

In Experiment 2, we asked participants to provide an explicit free-text guess about the nature of the causal relationship(s) being tested after they completed the learning phase. Eighty-six percent of these total responses (88/102) were compatible with the relevant learning observations, and here we only analyze these. Two independent coders categorized participants guesses according to their specificity and implicit localization of causal powers. The first coder categorized all free responses, and the categorized 15% were then compared against the first coder’s. Agreement level was 92%. The full set of free responses and the detailed coding scheme are available at https://github.com/bramleyccslab/causal_objects.

Since our ground truths are not the only rules consistent with the learning data, we analyzed participant self-reports not according to whether they got the ground truths right, but whether their own rules were consistent with the learning data, as well as the level of generality in the reports. Hence, we first defined three exclusive and exhaustive response specificity categories: *specific*, *fuzzy*, and *tacit*. A *specific* self-report would predict a unique result object for any potential combination of agent and recipient (for example “The inactive shape is always changed to a pentagon & its shade is changed to one step darker than the active stone”). A *fuzzy* rule was one that left open for more than one possible result objects (for example “It will be different colors and shapes”). We distinguished a second form of under-specified self-report, *tacit*, if it left a feature unmentioned, which depending on background assumptions might be taken to imply that feature remained unchanged but could also be compatible with it taking some new or random value (for example “The active stone adds a side to the inactive stone”).

We also had the coders categorize responses according to whether and how a self-report localized the domain of the causal law asserted. Concretely, we included four labels *A*, *R*, *AR*, and *universal*.

If a response mentioned a specific context of influence, typically using an *if...* clause, we labelled this according to whether the context mentioned the Agent (e.g. “If the active stone is darker than the inactive stone, it turns the inactive stone darker”), Recipient (e.g. “The active stone causes the other stones to change into a pentagon shape, unless it is already a pentagon shape, in which case it makes it darker”), or both. If a response made no localization or context (e.g. “The active stone cause inactive stones to five sided stone”) then it was labeled as *universal*.

Figure 8 illustrates the coding results by learning condition. Guess specificity is summarized in Figure 8A. We fit a multinomial logistic regression model predicting specificity by evidence-balance and ground truth factors, and found that when taking the *specific* self-report type as baseline, the ground truth factor is a significant predictor for the *tacit* type ($\beta = 1.54, p = 0.008$), while evidence-balance is not ($\beta = 0.78, p = \text{n.s.}$). Neither of these two factors are significant for the *fuzzy* type. Figure 8B summarizes participants’ guesses in terms of localization. No participant localized their rule in terms of both Agent and Recipient. Unsurprisingly, whenever localization occurred, it was applied with respect to the object that varied during the learning phase. A logistic regression predicting universal rule probability by condition showed that both evidence-balance (*fixed-recipient*, $\beta = -1.21, z = -2.3, p = 0.02$) and ground truth (*rule 2*, $\beta = 1.17, z = 2.3, p = 0.02$) were associated with more universal rules. There was no evidence for an interaction, $z = -0.5, p = \text{n.s.}$.

E.5 Model fits

We extended the grammar used in Experiment 1 to cover a larger space of ordered feature relationships. Concretely, we introduced +1, -1, >, < at the “bind relation” step to accommodate potential assertions about the ordering of feature values used in this experiment. As in Experiment 1, LoCaLa was expensive to evaluate so we optimised its parameters using a coarse grid search. Since there were six data points, during each iteration of the Gibbs sampler, when $\alpha = 5$ this observation has a half-half chance to create a new causal category or join the rest, in terms of category size preference, and this chance grows as α increases (Equation 8). Therefore, we centered the support values for α round 5, with an exponential increase for larger values, resulting in $\alpha \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 16, 32, 64, 128, 256\}$. β takes the same range of values as in fitting the models in Experiment 1. For γ , values of $\gamma = 1, 0.5$ and 0 are of particular theoretical interest, representing localization based on just the agent, agent and recipient equally, or just the recipient. We also included $\gamma = 0.25$ and $\gamma = 0.75$ consistent with a mixed focus biased toward either agent or recipient.