Beyond independent latent classes: Testing the limits of human flexibility

Pieter Verbeke[1], Maurice De Walsche[1]*, Pauline Maelfait[1]*, Tom Verguts[1]

* Both authors contributed equally to the manuscript

[1] Department of Experimental Psychology; Ghent University; Ghent; B9000

Abstract

A hallmark of human intelligence is the ability to flexibly adapt to novel situations. This flexibility relies crucially on the appropriate generalization of previously acquired knowledge. One influential theoretical framework argues that humans organize their knowledge in a collection of latent classes. Humans could then assign any novel situation to one of the latent classes (or construct a new one if it is too dissimilar), and thus generalize based on older knowledge. However, this framework is not sufficiently flexible to explain human generalization. In particular, we argue that at least three important features are missing: dependency, compositionality, and tuning. To empirically test this, we developed a novel behavioral task, which required adapting knowledge across tasks in order to generalize appropriately in the test phase. Across three experiments, dependency, compositionality, and tuning requirements were increasingly added to the basic task. Results demonstrate that humans are sensitive to all three types of structure. We discuss how current models must be extended to capture human generalization.

Keywords: Generalization, cognitive flexibility, learning, latent classes

Acknowledgements: This research was funded by research grant 3G010319 from Research Foundation Flanders. We thank Gaia Molinaro for helpful comments on a first draft of this manuscript. We thank Jonas Simoens for valuable input in the development and coding of the behavioral tasks.

Data and code availability: All data (https://osf.io/uynb3/) and code (https://github.com/CogComNeuroSci/PieterV_public/tree/master/Generalization_Behavioral) will be made available upon submission of the manuscript.

General introduction

Humans demonstrate a remarkable flexibility in dealing with novel challenges. A key aspect of this flexibility is to efficiently combine and adapt previously learned knowledge for the current task (Tomov et al., 2021; Vaidya & Badre, 2022; Verbeke & Verguts, 2024). For instance, while learning to drive a car is a complex learning process that requires practice, one can still partially rely on what was learned to navigate traffic as a cyclist or pedestrian (e.g., several traffic rules). This generalization across tasks significantly speeds up learning.

One way to transfer information efficiently between situations is to organize knowledge in a collection of latent classes, as formalized in the Dirichlet or Chinese restaurant process. Here, latent classes cluster together contexts, tasks or stimuli that share underlying structure. Originally applied in cognitive science to categorization (Anderson, 1991), and later to classical and operant conditioning (Gershman & Niv, 2012), this approach was more recently extended to clustering stimulus-action rules (tasks; Collins et al., 2014; Collins & Frank, 2013). In the context of stimulus-action rules, a latent class is often also referred to as task set since it clusters several tasks into a single class or set. For instance, one could create a latent class for "driving a car on the European mainland". This allows to generalize the stimulus-action sequences (i.e., the policy) that one learned in one car and one country to all cars and all countries on the European mainland. Different classes can be made for driving a bicycle (other action sequences) or driving a car in the UK (other traffic rules).

Critically, the latent class framework assumes that tasks are discretely assigned to one class from a collection of independent classes (Collins & Frank, 2013; Gershman & Niv, 2012; Razmi & Nassar, 2022; Vaidya et al., 2021). Information is then generalized to all tasks in the same class but not to other classes. In the current paper, we argue that this implementation can only account for a very narrow range of human generalization abilities. Below, we challenge three specific assumptions of the current latent class theory. Each of these challenges will be empirically tested in three different experiments.

A first assumption that we challenge is that latent classes are independent, with tasks being either fully correlated (when they are in the same class) or uncorrelated (when they are in a different class). Several real-world tasks exhibit a more complex relational structure. For example, the main difference between traffic rules in the UK and European mainland is simply that one should inverse the left-right dimension (driving left instead of right). Moreover, solving math problems often relies on understanding the anti-correlated nature of mathematical operations (e.g., addition versus subtraction). Independent latent classes do not allow such reasoning.

Second, we challenge the all-or-none nature of generalization. In the latent class framework, task rules are generalized to all tasks in the same class and not generalized when tasks are in different classes. However, policies for real-world tasks can rarely be transferred as a whole. For instance, while one can use a lot of knowledge from driving on the European mainland to driving in the UK, it remains critical to invert the left-right dimension for several traffic rules. To extend the latent class framework to such tasks, previous work proposed compositionality (Dekker et al., 2022; Franklin & Frank, 2018; Liu & Frank, 2022; Reverberi et al.,

2012). Here, task rules are divided in multiple components that are each assigned to a latent class. This effectively allows generalizing the action sequences of driving a car to the European mainland while avoiding interference from the UK traffic rules. Note that this approach still presumes independent classes, which would not allow to transfer the traffic rules in an inverted manner. This adds an empirically untested layer of complexity. Thus, we build on previous work proposing compositionality and extend this to anti-correlated compositional sets.

Third, we further elaborate on the all-or-none nature of generalization. Specifically, we challenge that (sub)sets of task rules are generalized in a rigid manner. Even when only generalizing a subset of the task rules, an adaptive agent may want to implement at least a minor tuning or transformation when generalizing to a different task (Verbeke & Verguts, 2024). For example, when transitioning from a small to a big car, most of the required action sequences are very similar. Nevertheless, a bigger car has a larger radius for making maneuvers, so some tuning of the learned action sequences is required.

Thus, we describe several arguments that challenge the current implementation of the latent class theory on human generalization (Collins et al., 2014; Collins & Frank, 2013; Gershman & Niv, 2012). Nevertheless, empirical investigations that test the limits of human flexibility in generalization are lacking. Therefore, we developed a novel paradigm and employed this in three different experiments, each one further increasing the complexity of task contexts. First, we challenge the notion of independence in the latent class framework. For this purpose, we test whether humans can generalize to anti-correlated task contexts (e.g., inverting the left-right dimension between UK and European mainland). A second experiment builds on previous work challenging the all-or-none nature of latent classes (Franklin & Frank, 2018, 2020; Liu & Frank, 2022). Specifically, we test compositional generalization. In contrast to previous work, this compositional generalization is tested in an environment combining both fully correlated and anti-correlated tasks, thus requiring subjects to decide what to simply transfer and what to invert. A third experiment further challenges the rigid nature of latent class generalization by testing more subtle tuning of task rules such as an expansion or shrinkage of the action space (comparable to switching from a small to big car).

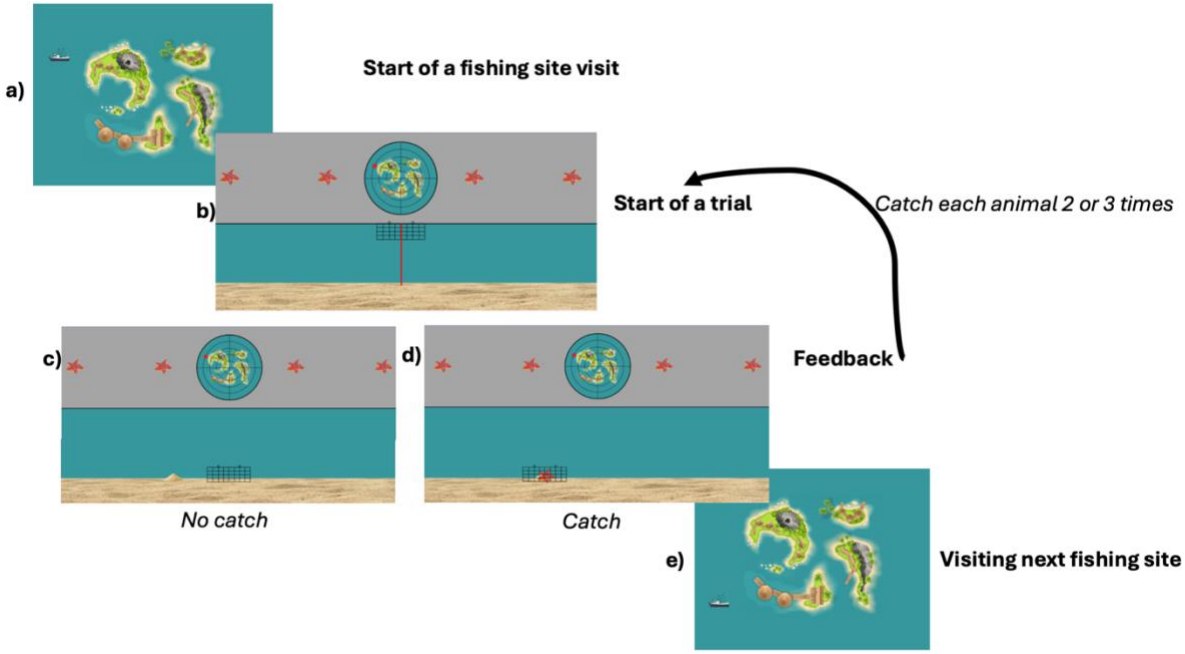
Experiment 1: Generalization to anti-correlated task contexts

In a first experiment, we challenge the notion of independent latent classes. More specifically, we test whether humans can spontaneously (without explicit training) generalize across anti-correlated tasks by learning about relationships between classes.

To test this, we developed a novel task paradigm, which will be adapted for the second and third experiment. The task is framed as a fishing contest and is visualised in Fig. 1. Here, a boat brings participants to a fishing site located around an island (Fig. 1a and Fig. 1e). When arriving at this fishing site, participants should try to catch an underwater animal (fish, squid, crab, ...). This can be done by positioning a cage (on a continuous left-to-right axis) and dropping it at the correct location (see Fig. 1b). After a couple of trials, the boat brings participants to another fishing site. The participants must thus learn the hiding spots of the animals. Each animal had a fixed hiding spot which depended on the fishing site. Critically, there was a

relational structure between fishing sites. While the hiding spots at some fishing sites were fully correlated (animals were hidden at the same location), they were uncorrelated or even anti-correlated (hiding spots were mirrored along the vertical axis) with other fishing sites. Inference about this relational structure allows to significantly speed up learning. A full overview of the hiding spots (on a left (-1) to right (+1) axis) is given in Fig. 2b.

Figure 1. The task.



Note. Overview of the task procedure. In each experimental round, participants visit multiple fishing sites. **a)** Each visit starts with the presentation of a boat moving to the relevant fishing site. Once the boat arrives at the fishing site it is presented at this location for a couple of seconds after which a trial is started. **b)** On each trial, participants can see the location of the fishing site on the radar, the animal that they need to catch (presented four times, at the top of the screen) and the cage. The cage is always initialized at the middle of the screen; a laser pointer indicates where it would land if it were dropped. **c-d)** After participants drop the cage, they receive feedback. Here, the animal would appear from underneath the sand. If the cage drops on the animal, it remains presented in the cage (d). If the cage does not catch the animal, the animal disappears, and a small heap of sand remains where it was hidden (c). Note that in the test rounds, participants do not get feedback. In this case, the cage will just drop on the sand but neither the animal or the heap are presented. This trial procedure is repeated until each animal is caught two or three times, depending on the experiment and round within the experiment. **e)** Then, the screen with the boat is presented again, which brings the participant to the next fishing site.

In all experiments, there were four types of experimental rounds which reflect a crossing of two animal sets (initial set and generalization set) and two types of feedback (learning rounds with feedback and testing rounds without feedback). In *initial learning rounds*, the participant needs to learn the hiding spot of four animals (initial set) for each fishing site. *Initial test* rounds investigate what participants have learned about this initial set. In these rounds, no feedback is given, meaning that the animal will not reveal itself after a cage drop. In *generalization training* rounds, a novel set of (generalization) animals is introduced. Participants again must learn the hiding spot of each animal at each fishing site. Importantly, only 2 fishing sites are visited during these generalization training rounds. These 2 fishing sites are referred to as trained sites.

Critically, if participants have correctly inferred the relationship between fishing sites during the initial learning rounds, the 2 trained sites provide sufficient information to know where the generalization animals are hidden at the other fishing sites. This is what is tested in the *generalization test* rounds. Here, participants visit all fishing sites and must try to catch animals from the generalization set introduced in the generalization training rounds. Again, no feedback is provided during these rounds. An extra overview of the difference between experimental rounds is provided in Table 1.

Table 1. Experimental rounds.

	Animal set	Feedback	Visited fishing sites
Initial learning round	Initial set	Yes	All
Initial test round	Initial set	No	All
Generalization learning round	Generalization set	Yes	Trained (2)
Generalization test round	Generalization set	No	All

Method

Participants. Fifty-three participants were recruited from the Sona participant pool at Ghent University. Two participants were excluded because they did not complete all experimental sessions. Technical problems caused data loss of another 2 participants. This resulted in a final sample of forty-nine participants (12 male, 37 female) with a mean age of 22.6 (complete age range: 18-32). Each participant had normal or corrected-to-normal vision. All participants gave their informed written consent before the experiment. To participate in three one-hour long sessions of which the last one was in the fMRI scanner, participants received a payment of 60 euros. The study was approved by the local ethics committee of Ghent University Hospital.

Apparatus and stimuli. The web-based experiment used in the first two sessions (procedure described below) was programmed using JsPsych (de Leeuw, 2015). For these sessions, participants could use their own pc if their screen had a standard 1.78 aspect ratio and 60 Hz refresh rate. Participants could position the cage by using the left and right arrow keys on their keyboard and drop the cage by pressing the spacebar. The task for the on-site data collection of the third session was programmed in PsychoPy (Peirce et al., 2019). Here, participants observed the screen via the mirror in the head coil and responded via a Cedrus response box. The cage could be positioned by using 2 response buttons at their right hand and dropped by pressing a response button with their left hand. The visual stimuli (e.g., underwater animals) were freely available on vecteezy.com.

Procedure. Participants had to complete three sessions of about one hour each. The first two sessions were performed in the participant's home environment via a web-based experiment. In these sessions, initial training rounds were alternated with initial test rounds. In the initial training round, participants visited all fishing sites. During a visit, participants were first

presented with an island group and a boat (see Fig. 1a). After one second, this boat would move around the island group until it reached the relevant fishing site. The boat remained at this position for 2 seconds. Then, another screen is presented (Fig. 1b). Here, a radar at the top of the screen indicates at which fishing site the participant currently is. Additionally, four instances of an underwater animal are shown at the top of the screen (2 on each side of the radar) to inform participants which animal they need to catch. Additionally, a cage was on every trial initialised at the middle of the screen (just below the radar). This cage spanned 10% of the screen width. Importantly, participants could never catch an animal by dropping it at its initialized (middle) location. Participants used button presses to move the cage. A single button press would shift the cage for 3% of the screen width. Alternatively, participants could keep the response button pressed to move the cage faster to the left or right. A different button press was needed to drop the cage and catch the animal. If the response limit (6 seconds in Experiment 1, and 4 seconds in Experiment 2 and 3) was exceeded, the cage would automatically be dropped at its current position. Once the cage was dropped, an animal would appear from beneath the sand, providing feedback about its hiding spot. It took the cage 250 milliseconds to drop to the bottom. During this time, the animal was already presented at its hiding spot. If the animal was caught (Fig. 1c), it would be presented in the cage for another 750 milliseconds. If the cage missed the animal (Fig. 1d), the animal would disappear, and the empty cage was presented for 750 milliseconds together with a small heap of sand where the animal was hidden. Then, the next trial would start, meaning that the cage was initialized again at the middle of the screen and a different animal had to be caught. During each visit of a fishing site, all four animals had to be caught three times. After this, the boat would appear again and bring participants to the next fishing site (Fig. 1e). Once all fishing sites were visited (in random order) participants could take a short break. This was repeated for three times after which the initial test round was introduced. Here, participants again visited all fishing sites in random order. In each visit, all four animals had to be caught two times. Critically, they did not receive trial-by-trial feedback anymore. Here, the cage would drop but neither the animal nor the heap of sand would appear. They were informed about the total number of caught animals at the end of the round. This sequence of initial learning round and initial test round was repeated twice in each session.

The third session was performed at Ghent University Hospital while participants were lying in the MRI scanner. Current manuscript only discusses behavioral results. Here, participants were first provided a short reminder of what was learned before by means of another initial training round. Again, all fishing sites were visited. In each visit, all four animals had to be caught two times. Each fishing site was visited only once. Then, participants performed a long generalization training round. This round introduced a generalization set of three novel animals and visited only the two trained fishing sites. During each visit, participants needed to catch all three generalization animals three times. Both trained sites were visited 6 times in an alternating fashion. After this, a generalization test round was introduced. Here, participants were tested on the novel set of three animals. However, in these rounds the participants visited all fishing sites without receiving trial-by-trial feedback. At the end of each round, participants were always informed about the total number of animals they caught. In Experiment 1, the generalization test

rounds followed a specific sequence of fishing site visits. Here, each time one of the trained sites was visited after which one of the associated (fully correlated or anti-correlated) fishing sites were visited. As a result, trained sites were visited twice as often in these generalization test rounds. On each visit, each animal had to be caught twice. In total there were six generalization test rounds, but these were intermixed with four generalization training rounds (one visit of both trained sites) and two repetitions of the initial learning round (one visit of all fishing sites). An overview of the sequence of rounds in all sessions (and the performance in each round) is given in Fig. 2c.

Analyses. To evaluate performance, we computed a baselined error score.

$$Baselined\ Errorscore_{p,t} = \frac{\left|Cage\ drop\ location_{p,t} - \ Animal\ hiding\ spot_{p,t}\right|}{\sum_{r=1}^{1000}\left|R_{\ p,t,r} - \ Animal\ hiding\ spot_{p,t}\right|/1000} \quad (1)$$

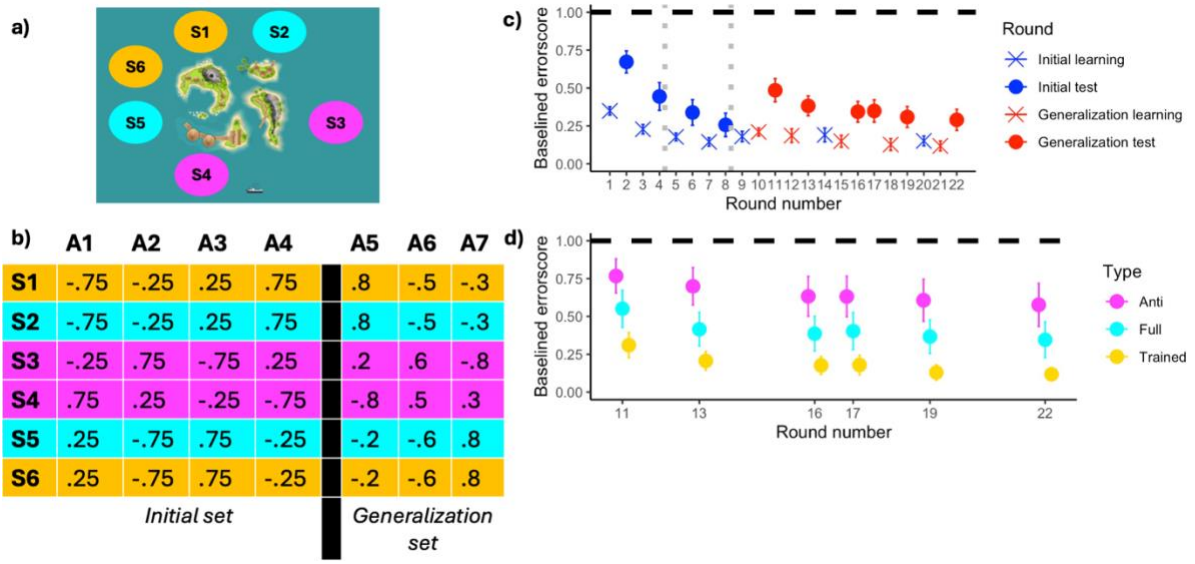
Here, on each trial, we compute the error score by taking the absolute distance (in screen width %) between the location where the participant (*p*) dropped the cage and the hiding spot of the animal on that trial (*t*). As baseline (denominator of Equation (1)), we take the average of 1000 replications of a random agent. On each replication (*r*), a random value $R_{p,t,r}$ is drawn from a uniform distribution between 0 and 1. Also hiding spots and cage drop locations are recoded on a 0 to 1 axis. Similar to the cage drop location, this random value is compared to the animal hiding spot on that trial. The average of 1000 replications was considered to reflect the error score of a random agent and is used to baseline the error score of the participant (see Equation (1)). As a result, the baselined error score gives an indication of performance where 0 is perfect performance (no distance between cage drop and hiding spot) and a value of 1 or more reflects random or worse than random performance.

To provide an overview of performance, we used a repeated measures ANOVA with this baselined error score as dependent variable and round number, feedback (test vs learning round) and animal set (initial versus generalization) as independent variables.

However, our main analyses are focused on performance in the generalization test rounds. Therefore, we also performed a repeated measures ANOVA on this subset of the data with the baselined error score as dependent variable. The independent variables were round number and type of fishing site (trained, fully correlated, or anti-correlated).

Results

Figure 2. Experiment 1.



Note. **a)** Overview of how fishing sites were organized around the island group. Different rotated versions of this organization were counterbalanced across participants. Color codes are also used in b and d and reflect the type of fishing site (see legend in d). **b)** provides an overview of the hiding spots of the animals (A1-A7) in all fishing sites (S1-S6). Values represent the hiding spot on a -1 (left) to 1 (right) range but this was linearly converted to a 0 to 1 range to compute error scores. Animals 1 to 4 belong to the initial set and animals 5-7 belong to the generalization set. **c)** illustrates the mean performance (as baselined error score; see Equation (1)) for each experimental round. Error bars provide 95% confidence intervals. The vertical grey dotted lines separate different sessions of the experiment. The horizontal black dashed line reflects error scores for a random agent. **d)** zooms in on the performance in different types of fishing sites during the generalization test rounds.

We first describe general performance analyses across all experimental rounds. These revealed a main effect of round number ($F(1,48) = 55.68$, $p < .0001$, $\eta^2 = .54$), indicating that the error scores significantly decreased over rounds (see Fig. 2c). Also the effect of feedback was significant ($F(1,48) = 114.5$, $p < .0001$, $\eta^2 = .70$), showing that error scores were significantly higher in the test rounds (no feedback; $M = .387$) than training rounds (with feedback; $M = .184$). Additionally, the effect of animal set (initial or generalization) also reached significance ($F(1,48) = 7.036$, $p = .011$, $\eta^2 = .13$). This revealed that on average the error scores were lower for the generalization animals ($M = .268$) than for the initial animals ($M = .285$). Also the interactions between round number and feedback ($F(1,48) = 12.58$, $p = .0008$, $\eta^2 = .21$) and between feedback and animals ($F(1,48) = 47.97$, $p < .0001$, $\eta^2 = .50$) reached significance. The interaction between round number and animals did not reach significance ($F(1,48) = .032$, $p = .859$, $\eta^2 < .0001$). The three-way interaction between round number, feedback and animals did reach significance ($F(1,48) = 63.33$, $p < .0001$, $\eta^2 = .57$). Together, these interactions indicates that the difference between test and training performance decreased over rounds and that this decrease was stronger for the initial animals than for the generalization animals.

We next investigated whether participants could generalize to full and anti-correlated fishing sites. For this purpose, we zoomed in on performance during the generalization test rounds (Figure 2d). Analyses of the performance (baselined error score; see Equation (1)) in the generalization test rounds, revealed again a main effect of round number ($F(1,48) = 3.045$, $p < .0001$, $\eta^2 = .36$), indicating that the error score decreased over rounds (Fig. 2d). Notice that

participants never received feedback in these rounds but did receive feedback in the round(s) in between two generalization test rounds. Hence, this learning effect reflects transfer from what was learned in between two generalization test rounds. Also the effect of fishing site type reached significance ($F(2, 96) = 31.82$, $p < .0001$, $\eta^2 = .45$). Here, paired t-tests revealed a clear order in performance. Specifically, the baselined error score was lowest in the trained sites ($M = .187$, $SD = .162$). This was significantly better ($t(48) = 5.668$, $p < .0001$, $d = .81$) than the fully correlated fishing sites ($M = .411$, $SD = .343$). This was again significantly better ($t(48) = 3.982$, $p = .0002$, $d = .57$) than the baselined error score in the anti-correlated fishing sites ($M = .652$, $SD = .401$). Importantly, performance was significantly better than random (i.e., baselined error score smaller than 1) in all three types of fishing sites (all $p < .0001$). The interaction between round number and fishing site type did not reach significance ($F(2, 96) = .025$, $p = .976$, $\eta^2 < .0001$), revealing that the order in performance across types of fishing sites was stable across rounds.

Discussion

Experiment 1 demonstrates that humans can learn about more complex task relations than the latent class framework allows. Their generalization was much better than random in the fully correlated fishing sites (as predicted by the latent class framework), but also in the anti-correlated fishing sites (which is not predicted by this framework). Presumably, extracting not only the latent classes but also their relational structure allowed them to generalize efficiently.

Previous experimental work only tested generalization in fully correlated task contexts (Collins et al., 2014; Collins & Frank, 2013; Vaidya et al., 2021). We replicated this finding but also went beyond previous work by introducing anti-correlated tasks. Here, participants had to invert the mappings that they learned in one of the trained tasks. Although performance was significantly worse in the anti-correlated tasks compared to the fully correlated ones, performance within the anti-correlated tasks was still significantly better than random. Hence, human generalization can profit from dependence between latent classes.

Experiment 2: Compositional generalization in anti-correlated task contexts

In a second experiment, we challenge the all-or-none nature of human generalization. Consistently, previous work proposed a compositional extension of the latent class framework (Franklin & Frank, 2018; Liu & Frank, 2022). This allows to transfer part of the task rules from one context to another while not generalizing other parts of the task rules. In Experiment 2, we explored a more complex task environments in which fully and anti-correlated transfers interacted with compositional generalization.

Specifically, we adapted the experimental paradigm from Experiment 1 but added two compositional fishing sites. In one of the two compositional fishing sites, the hiding spots of the outer two animals (out of four) were inverted compared to the trained sites and the inner two animals had the same hiding spot as in the trained sites (e.g., S5 relative to S1; see Fig. 3b). In the second compositional fishing site, the hiding spots of the inner two animals were inverted while the hiding spots of the outer two animals were the same as in the trained sites (e.g., S4 relative to S1; see Fig. 3b). The other 6 fishing sites were the same as in Experiment 1. This allowed us to

replicate our findings. Fig. 3a and Fig. 3b provide an overview of the fishing sites (S1-S6) and the animal (A1-A8) hiding spots at each fishing site.

Method

Participants. Forty-five students from Ghent University were recruited in exchange for course credits. Four participants were excluded because they did not complete all experimental sessions. Additionally, four participants were removed before the analyses because they performed worse than the random baseline in the last initial test round. Note that the qualitative patterns did not change when we included these participants. This resulted in a final sample of thirty-seven participants (6 male, 31 female). Each participant had normal or corrected-to-normal vision. All participants gave their informed written consent before the experiment.

Apparatus and stimuli. The same visual stimuli as for Experiment 1 were used. Here, all sessions used a web-based experiment, which was programmed by using JsPsych (de Leeuw, 2015). Again, participants could use their own pc if their screen had a standard 1.78 aspect ratio and 60 Hz refresh rate. Participants could position the cage by using the left and right arrow keys on their keyboard and drop the cage by using the spacebar.

Procedure. Participants had to complete two sessions of about one hour each. Similar as the first two sessions of Experiment 1, the first session of Experiment 2 implemented two alternations of an initial training round (each context visited three times with feedback) with an initial test round (each context visited once without feedback).

The second session consisted of three repetitions of the following sequence: initial learning round, generalization learning round, generalization test round. In initial learning rounds, participants visited all fishing sites. In each visit, all four animals had to be caught two times. In generalization learning rounds, a generalization set of four novel animals was used. Here, only the two trained sites were visited. Each of these fishing sites were visited three times in alternating order. On each visit, all four generalization animals needed to be caught two times. In the generalization test rounds, participants visited each fishing site once. Also here, all four generalization animals had to be caught two times during each visit. Critically, as in Experiment 1, participants received no feedback during these generalization test rounds.

Analyses. To evaluate performance, we again computed the baselined error score (Equation (1)). To provide an overview of general performance, we used a repeated measures ANOVA with this baselined error score as dependent variable and round number, feedback and animal set as independent variables.

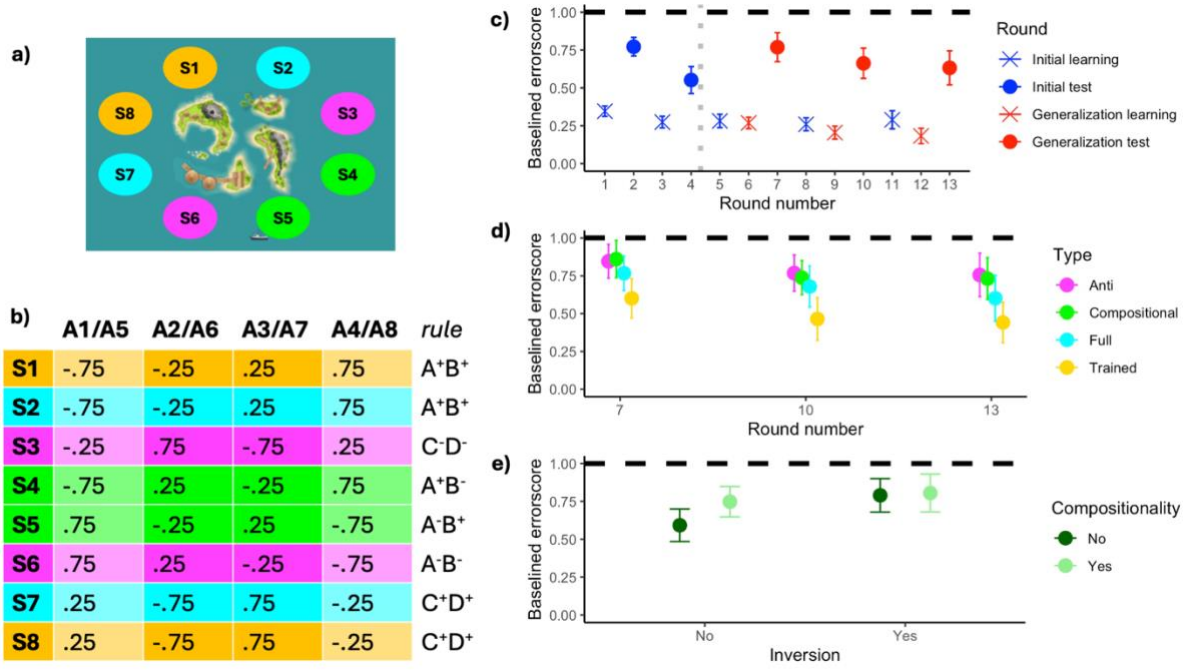
Similar to Experiment 1, our main analyses are focused on performance in the generalization test rounds. Here, we used a repeated measures ANOVA with the baselined error score as dependent variable. Again, the independent variables were round number and fishing site type. However, in Experiment 2, there were four types of fishing sites: trained, fully correlated, anti-correlated and compositional.

Additionally, we performed more detailed analyses to explore a possible interaction between anti-correlated and compositional generalization. For this purpose, we decomposed each fishing site in 2 task rules (see shading in Fig. 3b). More specifically, the hiding spots of the

outer animals of the first trained site were coded as rule A^+ and the hiding spot of the inner animals as rule B^+. For the second trained site, the hiding spots of the outer animals were coded as rule C^+ and the hiding spots for the inner animals as rule D^+. As a result, fully correlated fishing sites followed rules A^+B^+ or C^+D^+ and anti-correlated fishing sites followed A^-B^- and C^-D^-. The two compositional sites followed task rules A^-B^+ and A^+B^- (for a full overview see Fig. 3b). We computed the baselined error score for each task rule in each fishing site. This score was then used as dependent variable in a repeated measure ANOVA which included two independent variables. A first factor was inversion (yes for anti-correlated ($^-$) task rules and no for fully correlated ($^+$) task rules). The second factor was compositionality (yes for compositional fishing sites and no for fully- and anti-correlated fishing sites). The performance in the trained contexts was omitted for these analyses.

Results

Figure 3. Experiment 2.



	A1/A5	A2/A6	A3/A7	A4/A8	*rule*
S1	-.75	-.25	.25	.75	A^+B^+
S2	-.75	-.25	.25	.75	A^+B^+
S3	-.25	.75	-.75	.25	C^-D^-
S4	-.75	.25	-.25	.75	A^+B^-
S5	.75	-.25	.25	-.75	A^-B^+
S6	.75	.25	-.25	-.75	A^-B^-
S7	.25	-.75	.75	-.25	C^+D^+
S8	.25	-.75	.75	-.25	C^+D^+

Note. **a)** presents an overview of how fishing sites were organized around the island group. Different rotated versions of this organization were counterbalanced across participants. The color codes are also used in b and d and reflect the context type (see legend in d). **b)** provides an overview of the hiding spots of the animals (A1-A8) in all fishing sites (S1-S8). Animals 1-4 belong to the initial set and animals 5-8 belong to the generalization set. Animals were also decomposed in two rule sets. This is reflected by the saturation of the color shading. As a result, each fishing site followed two task rules (see right column). Values in the table represent the hiding spot on a -1 (left) to 1 (right) range but this was converted to a 0 to 1 range to compute error scores. **c)** illustrates the mean performance (as baselined error score; see Equation (1)) for each experimental round. Error bars provide 95% confidence intervals. The vertical grey dotted lines separate different sessions of the experiment. The horizontal black dashed line reflects error scores for a random agent. **d)** zooms in on the performance in different types of fishing sites during the generalization test rounds. **e)** illustrates the interaction between compositionality and inversion. Specifically, mean baselined error scores are presented based on whether the task rule required inversion (-) or not (+) and whether the fishing site was

compositional or not (full or anti; trained was omitted here). Again, error bars provide 95% confidence intervals and the horizontal black dashed line reflects error scores for a random agent.

General performance analyses again revealed a main effect of experimental round ($F(1,36) = 6.33$, $p = .0165$, $\eta^2 = .15$). Indeed, Fig. 3c illustrates that the error scores significantly decreased over rounds. Also the effect of feedback was significant ($F(1,36) = 223.5$, $p < .0001$, $\eta^2 = .86$), showing that error scores were significantly higher in the test rounds ($M = .677$) than training rounds ($M = .263$). The effect of animals (initial versus generalization) did not reach significance in Experiment 2 ($F(1,36) = 0.098$, $p = .756$, $\eta^2 < .0001$). In contrast to Experiment 1, the interaction between round number and feedback ($F(1,36) = 1.215$, $p = .278$, $\eta^2 = .03$) did not reach significance. However, the interaction between round number and animals ($F(1,36) = 11.38$, $p = .002$, $\eta^2 = .24$), as well as the interaction between feedback and animals ($F(1,36) = 31.02$, $p < .0001$, $\eta^2 = .46$) reached significance. Additionally, the three-way interaction between round number, feedback and animals did reach significance ($F(1,36) = 23.93$, $p < .0001$, $\eta^2 = .40$). Thus, the difference between the initial and generalization animals decreased over rounds. This decrease was mainly driven by the generalization learning rounds.

More detailed analyses on the generalization test rounds (see Fig. 3d), also demonstrated a main effect of round number ($F(1,36) = 13.06$, $p = .0009$, $\eta^2 = .27$). More importantly however, the effect of fishing site type reached significance ($F(3, 108) = 16.15$, $p < .0001$, $\eta^2 = .31$). Here, paired t-tests revealed the same order in performance as for Experiment 1. Specifically, the baselined error score was lowest in the trained sites ($M = .503$, $SD = .352$). This was significantly better ($t(36) = 4.329$, $p = .0001$, $d = .71$) than the fully correlated fishing sites ($M = .683$, $SD = .341$). This was again, significantly better ($t(36) = 2.501$, $p = .017$, $d = .41$) than the baselined error score in the anti-correlated fishing sites ($M = .792$, $SD = .331$). Performance in the two compositional sites ($M = .775$, $SD = .307$) did differ significantly from the trained sites ($t(36) = 5.469$, $p < .0001$, $d = .9$), but did not show significant differences with the fully correlated sites ($t(36) = 1.909$, $p = .064$, $d = .31$) or the anti-correlated fishing sites ($t(36) = .355$, $p = .725$, $d = .06$). Note that performance was more comparable to the anti-correlated fishing sites. Importantly, performance was significantly better than random (i.e., baselined error score smaller than 1) in all four context types (all $p < .0003$). The interaction between round number and context type did not reach significance ($F(3, 108) = .302$, $p = .824$, $\eta^2 < .0001$).

We next investigated an interaction between compositionality and the nature of task transformations (inversion (in anti-correlated rules) vs no inversion (in fully correlated rules)). Here, we find a strong effect of inversion ($F(1,36) = 13.59$, $p = .0007$, $\eta^2 = .27$) but also a significant effect of compositionality ($F(1,36) = 4.899$, $p = .033$, $\eta^2 = .12$). Interestingly, also the interaction between inversion and compositionality reached significance ($F(1,36) = 5.703$, $p = .022$, $\eta^2 = .14$). As shown in Fig. 3e, these results indicate that performance is worse when inversion (anti-correlation) is needed. Performance is also worse in compositional sites but only for the fully correlated task rule. There is no added effect of compositionality for the anti-correlated task rules.

Discussion

Experiment 2 replicated the findings of Experiment 1. Again, we found generalization across fully- and anti-correlated task contexts. As in Experiment 1, generalization was better for the fully correlated fishing sites than for the anti-correlated fishing sites.

Furthermore, we found that also when compositional generalization is required, participants can exploit anti-correlations between task rules. In general, both compositional and anti-correlated generalization tended to be more difficult than generalization across fully correlated task rules. However, there appears to be no additive difficulty effect when compositionality and anti-correlation are combined.

Experiment 3: Testing the limits of task rule tuning

The previous experiments suggest that human generalization is more flexible than the all-or-none process described in the latent class framework. In Experiment 3, we aimed to further test the limits of these generalization abilities. Specifically, we test whether humans can also learn about, and tune learned task rules, such as would be required to switch from a big to a small car and vice versa.

For this purpose, we add two novel types of fishing sites. One fishing site investigates expansion of the stimulus space while the other fishing site investigates shrinkage of the stimulus space. In total, there were 8 fishing sites (Fig. 4a). Again, two fishing sites function as trained sites. Two other fishing sites are fully correlated with one of these trained sites. A fifth fishing site is anti-correlated to one of the trained sites. A sixth fishing site substitutes the hiding spots of two animals. This is similar to one of the compositional sites in Experiment 2. The seventh fishing site expands the hiding spots in the trained sites (i.e., putting the animals further apart). The eighth fishing site shrinks the hiding spots in the trained site (i.e., putting the animals closer together). Critically, even in the shrinkage condition, no animal can be caught without moving the cage. An overview of the animal hiding spots at each fishing site is provided in Fig. 4b.

Method

Participants. Forty-five students from Ghent University were recruited in exchange for course credits. Five participants were excluded because they did not complete all experimental sessions. Additionally, one participant was omitted from the analyses because they performed worse than baseline in the last initial test round. This resulted in a final sample of thirty-nine participants (2 Male, 37 Female). Each participant had normal or corrected-to-normal vision. All participants gave their informed written consent before the experiment.

Apparatus and stimuli. The same visual stimuli as for Experiment 1 and 2 were used. Here, all sessions used a web-based experiment, which was programmed by using JsPsych (de Leeuw, 2015). We used the same requirements and response buttons as in Experiment 2.

Procedure. The procedure was exactly the same as in Experiment 2. Participants had to complete two sessions of about one hour each. In the first session, initial training rounds were intermixed with initial test rounds. The second session consisted of three repetitions of the following sequence: initial learning round, generalization learning round, generalization test round.

Analyses. To evaluate performance, we used the baselined error score (Equation (1)). We again performed general performance analyses in the form of a repeated measures ANOVA with the baselined error score as dependent variable and round number, feedback and animal set as independent variables.

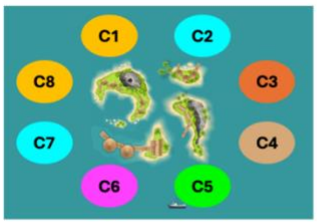
Similar to Experiment 1 and 2, our main analyses focused on performance in the generalization test rounds. Here, we performed a repeated measures ANOVA with the baselined error score as dependent variable. The independent variables were round number and type of fishing site (trained, fully correlated, anti-correlated, compositional, expansion and shrinkage).

Additional analyses were done to test whether participants performed an expansion or shrinkage in the two novel contexts and not simply used a fully correlated transfer from the trained context. For this purpose, we computed a confusion score. Specifically, we computed the absolute distance between the cage drop location and the hiding spot of the animal in the trained site. This reflects what the (unbaselined) error score would be if the participant performed on the trained site. We compared this confusion score to the unbaselined error score, which is just the absolute distance between the cage drop location and the hiding spot of the animal in the current (expansion or inversion) fishing site. If the unbaselined error score is lower than the confusion score, this indicates that participants tuned their behavior in the correct manner.

Results

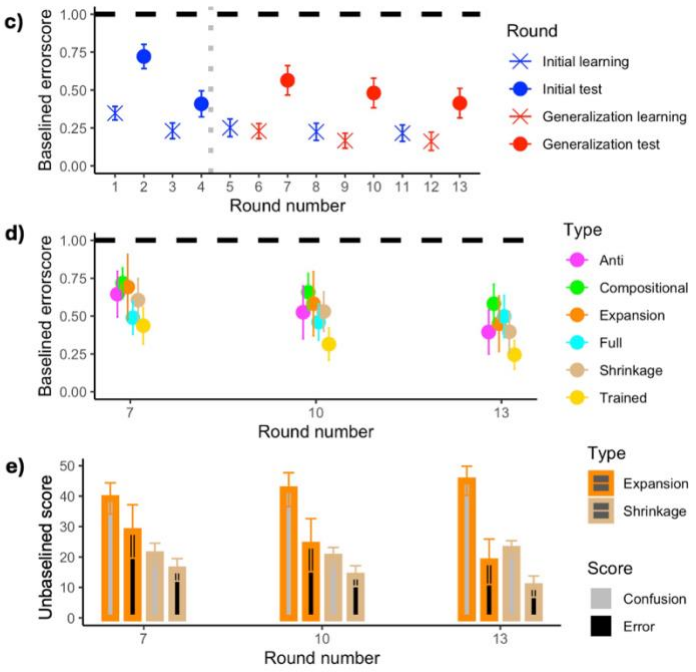
Figure 4. Experiment 3.

a)



b)

	A1/A5	A2/A6	A3/A7	A4/A8
S1	-.5	-.5	.5	.5
S2	-.5	-.5	.5	.5
S3	-.8	-.8	.8	.8
S4	-.2	-.2	.2	.2
S5	-.5	.5	-.5	.5
S6	.5	.5	-.5	-.5
S7	-.45	-.15	.15	.45
S8	-.45	-.15	.15	.45



Note. **a)** presents an overview of how fishing sites were organized around the island group. Different rotated versions of this organization were counterbalanced across participants. Color codes are also used in b and d and reflect the type of fishing site (see legend in d). **b)** provides an overview of the hiding spots of the animals (A1-A8) in all fishing sites (S1-S8). Animals 1-4 belong to the initial set and animals 5-8 belong to the generalization set. Values represent the hiding spot on a -1 (left) to 1 (right) range but this was converted to a 0 to 1 range to compute error scores. **c)**

illustrates the mean performance (as baselined error score; see Equation (1)) for each experimental round. Error bars provide 95% confidence intervals. The vertical grey dotted lines separate different sessions of the experiment. The horizontal black dashed line reflects error scores for a random agent. **d)** zooms in on the performance in different types of fishing sites during the generalization test rounds. **e)** illustrates the confusion and unbaselined error scores in both the expansion and shrinkage sites for each generalization test round. This is given in screen width %. Error bars provide 95% confidence intervals.

General performance analyses (Fig. 4c) again revealed a learning effect across experimental rounds ($F(1,38) = 40.71$, $p < .0001$, $\eta^2 = .52$). Also the effect of feedback was significant ($F(1,38) = 115.7$, $p < .0001$, $\eta^2 = .75$), showing that error scores were significantly higher in the test rounds ($M = .517$) than training rounds ($M = .228$). The effect of animals (initial or generalization) did not reach significance in Experiment 3 ($F(1,38) = 0.006$, $p = .941$, $\eta^2 < .0001$). Again, the interaction between round number and feedback ($F(1,38) = 2.069$, $p = .159$, $\eta^2 = .05$) did not reach significance. Also the interaction between round number and animals ($F(1,38) = .197$, $p = .66$, $\eta^2 < .0001$) was not significant. However, the interaction between feedback and animals ($F(1,38) = 25.45$, $p < .0001$, $\eta^2 = .40$) as well as the three-way interaction ($F(1,38) = 39.27$, $p < .0001$, $\eta^2 = .51$) reached significance. Together, this illustrates that the difference between initial and generalization animals decreased over rounds, and this decrease was mainly driven by the generalization learning rounds.

When performing more detailed analyses on the data from the generalization test (Fig. 4d) rounds, we observed that, even in this very complex environment, performance was significantly better than random (i.e., baselined error score smaller than 1) in all fishing sites (all $p < .0001$). Consistent with the previous experiments, the ANOVA demonstrated a main effect of round number ($F(1,38) = 18.50$, $p = .0001$, $\eta^2 = .33$). Moreover, the effect of the type of fishing site reached significance ($F(5, 190) = 6.86$, $p < .0001$, $\eta^2 = .15$). Again, performance was best in the trained sites ($M = .332$, $SD = .275$). This was significantly better than all other fishing sites (all $p < .005$). In contrast to the previous experiments, there was no significant difference in performance between the fully correlated fishing sites and the anti-correlated fishing sites ($t(38) = 0.943$, $p = .352$, $d = .15$). Nevertheless, overall performance was still better in the fully correlated fishing site ($M = .483$, $SD = .328$) than in the anti-correlated site ($M = .522$, $SD = .348$). Notably, performance was worst in the compositional fishing site ($M = .653$, $SD = .292$). This was significantly worse than all other fishing sites (all $p < .006$) except the expansion site ($M = .575$, $SD = .570$). The interaction between round number and context type again did not reach significance ($F(5, 190) = 2.12$, $p = .065$, $\eta^2 = .05$).

As described in the Analyses section, we used a confusion score to evaluate whether participants adapted behavior from the trained contexts to perform the task in the Expansion and Shrinkage contexts. As shown in Fig. 4e, the confusion score in the shrinkage context ($M = 21.078$, $SD = 6.235$) was significantly ($t(38) = 3.856$, $p = .0004$, $d = .62$) higher (worse) than the unbaselined error score ($M = 13.264$, $SD = 8.709$). Also in the Expansion fishing site, the confusion score ($M = 42.127$, $SD = 14.441$) was significantly ($t(38) = 3.186$, $p = .003$, $d = .51$) higher than the unbaselined error score ($M = 23.565$, $SD = 23.349$). Hence, instead of simply transferring task rules, participants tuned behavior to make adaptive changes to the learned task rules.

Discussion

Experiment 3 went one step beyond Experiments 1 and 2 and investigated whether, compared to the all-or-none generalization process described in the latent classes framework, humans would also tune existing task rules to behave more adaptively. Results indeed suggest that participants did so. More generally, participants demonstrated a remarkable flexibility in learning about a wide variety of complex task relations and exploiting this for generalization.

Notably, we found that, compared to all other fishing site, the compositional sites were significantly harder to learn in Experiment 3. However, this is probably due to our design choices. Since we used a wide variety of fishing site types, we chose to limit the complexity of the learning process by reducing the number of unique hiding spots. As can be observed in Figure 4b, some animals were hidden at the same spot within one fishing site (both .5 or both -.5), which probably caused participants to cluster these two animals together during learning and made it harder to tear them apart again for generalization in the compositional fishing sites.

General Discussion

Across three experiments, we robustly demonstrate three features of human generalization that are incompatible with traditional latent class theory, namely dependency, compositionality and tuning. Each experiment added a novel test to the previous ones, so that our results also yield internal replications. We argue that latent class models should be made sufficiently expressive to capture human generalization.

First, we challenged the notion of independence between latent classes. We demonstrated that, on top of generalizing across fully correlated tasks, humans can also generalize across anti-correlated tasks. Interestingly, there was a significant difference in generalization performance between fully correlated and anti-correlated tasks. This suggests that there is an added difficulty in learning about anti-correlated task relations compared to fully correlated task relations. Such difficulties could be explained by assuming a hierarchical learning process in which learning about relationships between latent classes happens at a higher level of abstraction than the construction of latent classes itself (e.g., Kemp et al., 2010).

Second, we challenged the all-or-none nature of the latent class framework. Here, we built on previous work proposing compositionality (Franklin & Frank, 2018, 2020; Liu & Frank, 2022), but we combined this with the anti-correlations described in Experiment 1. Results indicated that also when compositional generalization is required, participants can exploit anti-correlations between task rules. Interestingly, both compositional and anti-correlated generalization tended to be more difficult than generalization across fully correlated task rules. Moreover, there appears to be no additive difficulty effect when compositionality and anti-correlation are combined. This suggests that both types of generalization share a similar level of abstraction, which is higher than the generalization across fully correlated task contexts.

In a third step, we elaborated on the all-or-none nature of generalization in the latent class framework. Here, we tested whether humans would be able to learn about, and apply, task rule tuning. Specifically, we added two fishing sites that required to either expand or shrink the

geometrical space of the hiding spots. Participants were able to generalize such expansion and shrinkage as well.

As proposed by the latent class framework, it is important to sample from existing (classes of) knowledge when encountering a novel challenge. Indeed, previous empirical work has demonstrated that people prioritize past solutions in novel situations (Hall-McMaster et al., 2024). Nevertheless, this is often just an optimal starting point for learning (Tomov et al., 2021). In most cases, considerable updates are still required to behave adaptively in the novel context. Moreover, novel situations often require combining knowledge from multiple classes that one learned before. As we mentioned before, a partial solution is compositionality, which allows to combine parts of information from different classes for generalization. However, also this solution is not sufficiently flexible to explain the tuning (inversion, shrinkage and expansion) that we tested in current work.

One conceptual extension that allows for dependence between classes is to use an Indian buffet process as prior in assigning stimuli to classes (Griffiths & Ghahramani, 2011) instead of the Chinese restaurant process (Griffiths et al., 2003). This approach removes the constraint that a collection of stimulus-action mappings (a task) can only belong to one class; this allows for compositional generalization in a more flexible manner than with traditional latent classes (Franklin & Frank, 2018). Nevertheless, also this type of model has difficulties in explaining why performance was consistently better for fully correlated than for anti-correlated fishing sites. Another extension of the latent class framework is to construct hierarchical classes (Franklin & Frank, 2020; Kemp et al., 2010; Liu & Frank, 2022). Here, latent classes would be clustered themselves in even higher order classes. As we briefly mentioned before, such an approach would allow to cluster fully correlated task contexts in one class and then on a more abstract level also represent anti-correlations between classes. However, to our knowledge no implementation of hierarchical classes exist that allow for anti-correlations between or within classes. Furthermore, it is unclear how one can implement classes that allow for subtle tuning as in our expansion and shrinkage task contexts.

In the latent class framework, each object (task in our case) is assigned to a single class. Other inference procedures have been proposed, such as inferring the values on several continuous latent dimensions for each object (as in variational Gaussian inference; Kingma & Welling, 2022), or assigning each object to several classes simultaneously (as in the Helmholtz machine; Dayan et al., 1995). Nevertheless, it is also not clear whether these approaches are sufficiently expressive to capture the rich structure that we observed. Moreover, they have rarely been used to explain human categorization or generalization across tasks.

A modelling framework with more expressivity than latent classes, are artificial neural networks. However, an often-articulated fear is that their representations are not sufficiently structured to capture human generalization (Fodor & Pylyshyn, 1988; Marcus, 2018). Interestingly, however, compositional representations have been shown to emerge naturally in neural networks that are trained to perform many tasks (Johnston & Fusi, 2023; Yang et al., 2019). Moreover, compositionality in neural networks emerges also for anti-correlated tasks (Yang et al., 2019). Other researchers even demonstrated that it allows to generalize in an anti-correlated

manner (Riveland & Pouget, 2024). Another approach that implements compositionality in a more explicit manner are mixture of expert networks. Here, an agent can adapt behavior by linearly combining different existing expert networks (Jacobs et al., 1991; Jordan & Jacobs, 1994). Hence, instead of an all-or-none process, a weighted integration of information from different experts can be used to create novel behavior. Interestingly, it has been proposed that the ventrolateral prefrontal cortex of the human brain would function as a gating region, integrating the weighted contributions of all experts in other brain areas (O'Doherty et al., 2021). Another interesting and relevant neural network concept is the adapter from Artificial Intelligence (Zhang et al., 2021). Here, extra processing layers are introduced in a pretrained network and only the extra processing layers are trained for a novel task. Such an approach could be particularly useful to capture the task-specific tunings that we observed in Experiment 3 (Lu et al., 2024).

Yet another way to organize tasks (or any other objects) are cognitive maps. Here, one maximizes the distance between orthogonal objects and clusters objects that share sufficient similarities (Behrens et al., 2018; Schuck et al., 2016; Wilson et al., 2014). Like the latent class framework, this approach will thus cluster similar task contexts together. However, it allows for more flexibility in the sense that the distance between two tasks can be parametrically modulated (allowing for more graded dependence between tasks), representations can be partially overlapping (not all-or-none), and other neural regions can still tune their influence on behavior. Previous work has proposed that cognitive maps are mainly present in human hippocampus, ventromedial prefrontal cortex and orbitofrontal cortex (Baram et al., 2021; Bernardi et al., 2020; Schuck et al., 2016; Vaidya et al., 2021; Wilson et al., 2014).

In sum, current work demonstrated that the flexibility of human generalization goes well beyond what can be captured by latent classes. Here, we specifically focused on dependency, compositionality, and tuning. To understand how humans generalize, a major challenge will be combining the structure of latent classes with the expressivity of neural networks.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M., & Behrens, T. E. J. (2021). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron*, *109*(4), 713–723. https://doi.org/10.1016/j.neuron.2020.11.024

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge

for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954–967. https://doi.org/10.1016/j.cell.2020.09.031

Collins, A. G. E., Cavanagh, J. F., & Frank, M. J. (2014). Human EEG Uncovers Latent Generalizable Rule Structure during Learning. *Journal of Neuroscience*, *34*(13), 4677–4685. https://doi.org/10.1523/JNEUROSCI.3900-13.2014

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. https://doi.org/10.1037/a0030852

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz Machine. *Neural Computation*, *7*(5), 889–904. https://doi.org/10.1162/neco.1995.7.5.889

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dekker, R. B., Otto, F., & Summerfield, C. (2022). Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences*, *119*(41). https://doi.org/10.1073/pnas.2205582119

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Franklin, N. T., & Frank, M. J. (2018). Compositional clustering in task structure learning. *PLoS Computational Biology*, *14*(4), 1–25. https://doi.org/10.1371/journal.pcbi.1006116

Franklin, N. T., & Frank, M. J. (2020). Generalizing to generalize: Humans flexibly switch between compositional and conjunctive structures during reinforcement learning. *PLoS Computational Biology*, *16*(4), 1–33. https://doi.org/10.1371/journal.pcbi.1007720

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning and Behavior*, *40*(3), 255–268. https://doi.org/10.3758/s13420-012-0080-8

Griffiths, T., & Ghahramani, Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185–1224.

Griffiths, T., Jordan, M., Tenenbaum, J., & Blei, D. (2003). Hierarchical Topic Models and the Nested Chinese Restaurant Process. *Advances in Neural Information Processing Systems*, *16*. https://proceedings.neurips.cc/paper_files/paper/2003/hash/7b41bfa5085806dfa24b8c9de0ce567f-Abstract.html

Hall-McMaster, S., Tomov, M. S., Gershman, S. J., & Schuck, N. W. (2024). *Neural Prioritisation of Past Solutions Supports Generalisation* (p. 2024.06.10.598294). bioRxiv. https://doi.org/10.1101/2024.06.10.598294

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*, *3*(1), 79–87. https://doi.org/10.1162/neco.1991.3.1.79

Johnston, W. J., & Fusi, S. (2023). Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-36583-0

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, *6*(2), 181–214. https://doi.org/10.1162/neco.1994.6.2.181

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to Learn Causal Models. *Cognitive Science*, *34*(7), 1185–1243. https://doi.org/10.1111/j.1551-6709.2010.01128.x

Kingma, D. P., & Welling, M. (2022). *Auto-Encoding Variational Bayes* (arXiv:1312.6114). arXiv. https://doi.org/10.48550/arXiv.1312.6114

Liu, R. G., & Frank, M. J. (2022). Hierarchical clustering optimizes the tradeoff between compositionality and expressivity of task structures for flexible reinforcement learning. *Artificial Intelligence*, *312*. https://doi.org/10.1016/j.artint.2022.103770

Lu, Q., Nguyen, T. T., Zhang, Q., Hasson, U., Griffiths, T. L., Zacks, J. M., Gershman, S. J., & Norman, K. A. (2024). *Reconciling Shared versus Context-Specific Information in a Neural Network Model of Latent Causes* (arXiv:2312.08519). arXiv. https://doi.org/10.48550/arXiv.2312.08519

Marcus, G. (2018). *Deep Learning: A Critical Appraisal* (arXiv:1801.00631). arXiv. https://doi.org/10.48550/arXiv.1801.00631

O'Doherty, J. P., Lee, S. W., Tadayonnejad, R., Cockburn, J., Iigaya, K., & Charpentier, C. J. (2021). Why and how the brain weights contributions from a mixture of experts. *Neuroscience & Biobehavioral Reviews*, *123*, 14–23. https://doi.org/10.1016/j.neubiorev.2020.10.022

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Razmi, N., & Nassar, M. R. (2022). Adaptive Learning through Temporal Dynamics of State Representation. *Journal of Neuroscience*, *42*(12), 2524–2538. https://doi.org/10.1523/JNEUROSCI.0387-21.2022

Reverberi, C., Görgen, K., & Haynes, J.-D. (2012). Compositionality of Rule Representations in Human Prefrontal Cortex. *Cerebral Cortex*, *22*(6), 1237–1246. https://doi.org/10.1093/cercor/bhr200

Riveland, R., & Pouget, A. (2024). Natural language instructions induce compositional generalization in networks of neurons. *Nature Neuroscience*, *27*(5), 988–999. https://doi.org/10.1038/s41593-024-01607-5

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, *91*(6), 1402–1412. https://doi.org/10.1016/j.neuron.2016.08.019

Tomov, M. S., Schulz, E., & Gershman, S. J. (2021). Multi-task reinforcement learning in humans. *Nature Human Behaviour*, *5*(6), Article 6. https://doi.org/10.1038/s41562-020-01035-y

Vaidya, A. R., & Badre, D. (2022). Abstract task representations for inference and control. *Trends in Cognitive Sciences*, *26*(6), 484–498. https://doi.org/10.1016/j.tics.2022.03.009

Vaidya, A. R., Jones, H. M., Castillo, J., & Badre, D. (2021). Neural representation of abstract task structure during generalization. *eLife*, *10:e63226.*, 1–22. https://doi.org/10.7554/eLife.63226

Verbeke, P., & Verguts, T. (2024). Reinforcement learning and meta-decision-making. *Current Opinion in Behavioral Sciences*, *57*, 101374. https://doi.org/10.1016/j.cobeha.2024.101374

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–279. https://doi.org/10.1016/j.neuron.2013.11.005

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X. J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306. https://doi.org/10.1038/s41593-018-0310-2

Zhang, R., Zheng, Y., Mao, X., & Huang, M. (2021). *Unsupervised Domain Adaptation with Adapter* (arXiv:2111.00667). arXiv. https://doi.org/10.48550/arXiv.2111.00667