Shared Explanatory Frameworks for Function Learning and Category Learning in Humans

Matt Wetzel

June 28, 2020

Contents

1 Introduction **1**

1.1 Function Learning 2

1.1.1 Models of Function Learning 6

1.2 Category Learning 9

1.2.1 Models of Category Learning 14

2 Key Areas of Present and Potential Overlap **15**

2.1 Prior Knowledge 15

2.2 Inference Learning 16

2.3 Dimensionality 18

2.4 Future Directions for Comparative Investigations 19

3 Further Discussion **23**

3.1 Flexibility of Monotonic Functions 23

3.2 Implications of Graphical Knowledge Representation 28

3.3 Relationship to Other Branches of Cognitive Science 33

3.4 Conclusion 34

1 Introduction

Our ability to successfully navigate through and understand our environment necessitates a sensitivity to the interrelationships between percepts, concepts, and events (both internal and external). For example, cooking a meal requires sensitivity to the functional relationships between temperature, quantity, and time, as well as the categorical relationships between ingredients and tastes. Much of the category learning (CL) literature has investigated the question of how categorical knowledge is acquired and generalized from and across percepts / concepts (Ashby and Maddox, 2005; K. J. Kurtz, 2015). Similarly, the function learning (FL) literature has focused on the question of how humans learn and extrapolate functional relationships between multiple continuously valued percepts / concepts (Busemeyer, Byun, Delosh, and McDaniel, 1997; Kalish, Lewandowsky, and Kruschke, 2004). Despite being seemingly disparate literatures, both involve mapping relationships between variables, and many of the leading theoretical explanations of FL and CL leverage the same computational frameworks (Busemeyer et al., 1997; Lucas, Griffiths, Williams, and Kalish, 2015). As such, the goals of the following paper are to (a) highlight the disparities and similarities between CL and FL, and (b) explore theoretical frameworks that provide an integrated explanation of both.

The next sections will briefly highlight the FL and CL literature, followed by a discussion of critical cross-domain phenomena that suggest mechanistic overlap. Finally, implications of integrative frameworks will be discussed.

1.1 Function Learning

Because variables[1] in our environment interact, learning functional relationships is crucial to an organism's ability to make inferences about present and future events. For example, adults should ideally have a keen awareness of the amount of alcohol they consume and the magnitude of the headache experienced the next day. The study of how people learn and generalize functional knowledge — a.k.a., function learning (FL) — has been a serious area of investigation in psychology warranted by the ubiquitousness of functional relationships in everyday life. In a typical FL experiment, subjects are presented with a continuously-valued stimulus (the **cue**); subjects then guess the value of a *second* feature (the **criterion**) given the value of the first (Busemeyer et al., 1997)[2]. Success is measured as the difference between a subject's guess (the **response**) and the value defined by the underlying function that generated the data. After training, subjects are shown new values of feature 1 that weren't observed during training (either inside or outside the training region); this allows researchers to observe how subjects interpolate and extrapolate — or, generalize — functional knowledge. Two primary phenomenological measures (among others) used by FL theorists to test explanatory models of human FL are (a) the difficulty through which some functions are learned relative to others, and (b) how subjects interpolate and extrapolate functional knowledge to new regions of space[3].

[4]The FL literature has been very productive in indexing the ease at which certain functions are learned, providing important benchmarks for theoretical models. Regarding *directionality*, increasing linear functions are easier to learn that decreasing linear functions

[1] both sensory and latent

[2] The types of features used in FL experiments vary, such as psychophysical cues (e.g., frequency; Koh and Meyer, 1991), spatial concepts (e.g., distance; Koh and Meyer, 1991), shape attributes (e.g., line length; DeLosh, Busemeyer, and McDaniel, 1997), abstract symbols (Koele, 1980), and even just numeric values themselves (Naylor and Clark, 1968).

[3] As will be discussed in a later section, FL and CL research are very, very similar in regards to their methodological paradigm and phenomenological measurements; though the use of absolute (CL) versus relative (FL) error measurements are a critical difference (as pointed out by Busemeyer et al., 1997).

[4] See Busemeyer et al., 1997 for a comprehensive review.

(Brehmer, 1974; Naylor and Clark, 1968). Regarding *linearity*, increasing linear functions are generally learned faster than increasing nonlinear functions (Byun, 1996; DeLosh et al., 1997). Regarding *monotonicity*, subjects have a harder time learning non-monotonic (e.g., the quadratic) than monotonic (e.g., the log, exponential, linear) functions (Brehmer, 1974; Byun, 1996; Carroll, 1963). Regarding *cyclicity*, non-monotonic cyclic functions (e.g., the sine) are learned slower than their non-monotonic, noncyclic counterparts (Bott and Heit, 2004; Byun, 1996)$^{\mathrm{I}}$[5]. Interestingly, some of the comparative differences in the adult learning literature might mirror the trajectory in which different functions are learned in childhood development (Ebersbach, Van Dooren, Van den Noortgate, and Resing, 2008; Ebersbach and Wilkening, 2007).

The comparative ease at which functions are learned is driven by more than just function family; there are a variety of variables regarding the presentation, framing, and duration of learning that impact both ease of learning and generalization. For example, difficult functions are learned faster when cues are presented in a monotonic sequence (Byun, 1996; DeLosh, 1995); possibly suggesting an important role of memory and recurrency during FL. Additionally, properties of the cues (or variables) use in FL studies are influenced by prior knowledge and expectations (Byun, 1996; Koele, 1980; Miller, 1971). For example, Koele (1980) found that subjects were better at recognizing positive relationships when using "meaningful" variable names that subjects might expect to be positively correlated, such as *intelligence* and *test scores* (relative to symbolic labels, such as X and Y). Another interesting phenomena is the bias towards linear representations at the early phases of learning nonlinear functions — as was found by Summers, Summers, and Karkau (1969) (though see Koh and Meyer, 1991). There also seem to be considerable subject-level variability in accuracy during function learning tasks (DeLosh et al., 1997; McDaniel,

[5] Note that all of these discussed findings involve the comparison of continuous functions, which are easier to learn than arbitrary, categorical functions (Carroll, 1963; Sniezek and Naylor, 1978).

Cahill, Robbins, and Wiener, 2014). Like in CL research, predicting the distribution of subject behaviors (as opposed to mean profiles) in an experiment is an interesting theoretical challenge for explanatory models of FL (as noted by Kalish et al., 2004; Lucas et al., 2015).

How subjects *generalize* functional knowledge also serves as a critical phenomenological metric in FL studies. Typically, generalization is distinguished into two types: **interpolation** (within the training region) and **extrapolation** (outside the training region). Subjects are generally much more accurate when interpolating functional knowledge relative to extrapolating it (DeLosh et al., 1997). Importantly, subjects often extrapolate in a way that's *inconsistent* with the original function learned (see figure 1), and instead tend to extrapolate in a manner best described by a much simpler function (specifically, in the same direction as the training function). For example, DeLosh et al. (1997) trained subjects on the center regions of linear, exponential, and quadratic functions — followed by a generalization phase than spanned beyond the space of the training cues. In addition to finding that mean generalization behavior deviated from the basis function in particular ways, DeLosh et al. (1997) highlighted qualitatively different learners that deviated from the averaged sample. For instance, when learning a quadratic function, some learners extrapolated the data in a manner best described by a linear or exponential function; other learners generalized the function perfectly along the quadratic (figure 1)[6]. These results cast doubts on prior theories that suggested humans learn functions by tuning internal representations of underlying basis functions[7] (Brehmer, 1974; Carroll, 1963).

[6] One difficulty in interpreting DeLosh et al. (1997)'s results is that the cue and response magnitudes were represented as horizontal bars that started at length 0; the particular response patterns elicited by humans may have been due to nonlinearities in the perception of the bar at different lengths (as originally suggested by Koh and Meyer (1991)).

[7] which is arguably a similar assumption made by connectionist models that rely on complex basis functions in the hidden layers

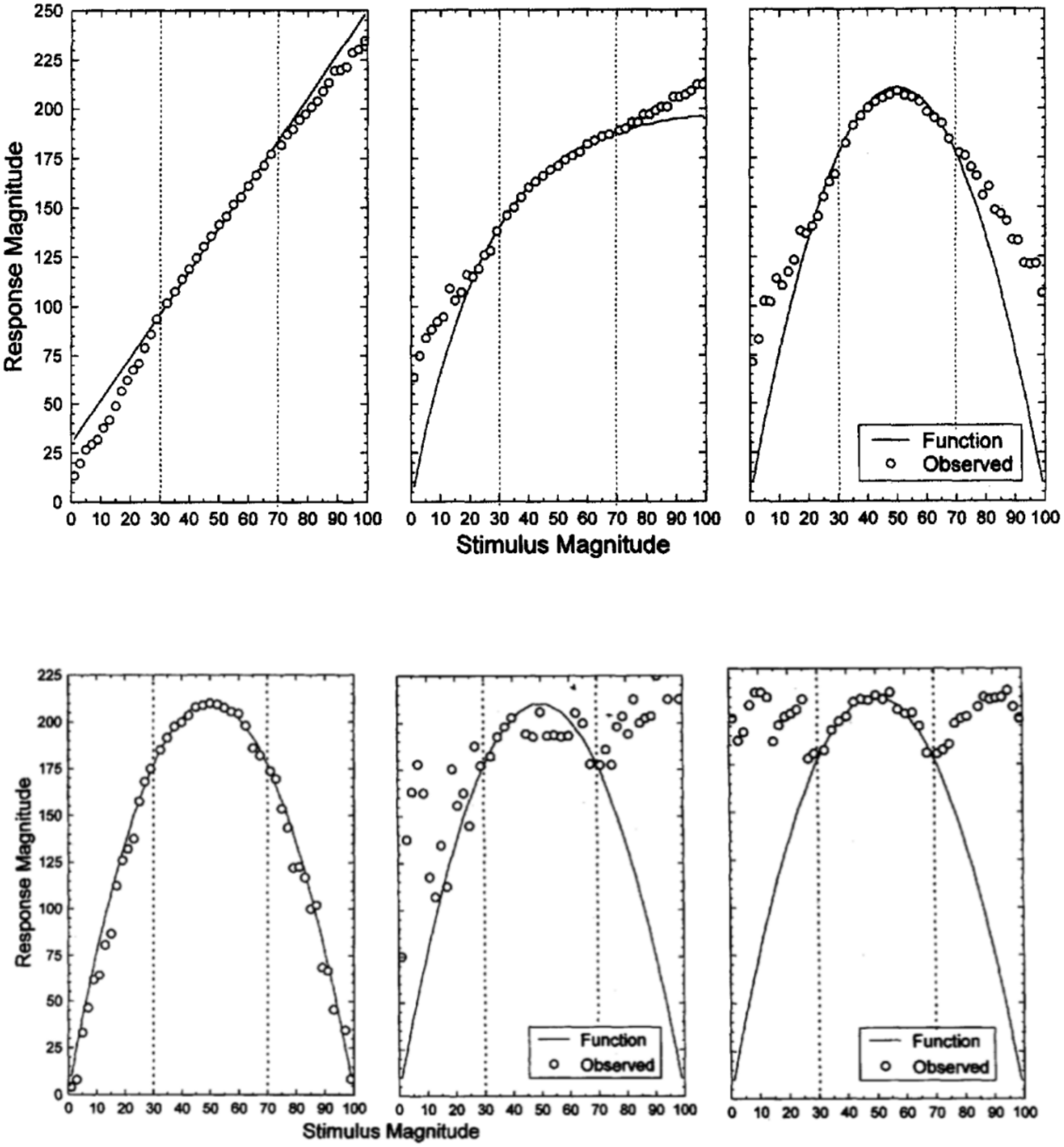

Figure 1: **Top**: Some of the mean extrapolation data from DeLosh, Busemeyer, and McDaniel (1997). Notice that the interpolation region is generalized much more accurately than the extrapolation region. **Bottom**: Examples of extrapolation data from 3 different subjects from the quadratic condition: one who learned the function almost perfectly, another that seemed to learn something akin to a linear function, and another who seemed to learn nothing at all. Taken directly from DeLosh, Busemeyer, and McDaniel (1997).

1.1.1 Models of Function Learning

Early models of function learning (referred to in the FL literature as “rule-based”) were analogous to statistical regression, assuming that psychological representations consisted of compositional basis functions that were tuned or selected for during learning (Brehmer, 1974; Koh and Meyer, 1991). For example, Brehmer (1974) found that interpolation behavioral data for a power, log, and linear function was best fit by a polynomial regression model biased towards power functions (relative to other regression models or associative stimulus-generalization models). While useful for predicting interpolation data, DeLosh et al. (1997)'s findings that humans often fail to extrapolate functional knowledge in a way that's consistent with the basis function presents a serious challenge for regression-based theories[II].

The phenomena that humans extrapolate functional knowledge in a manner *close-to-consistent* with the basis function also presents a serious problem for stimulus-generalization based theories (referred to in the FL literature as “associative”). Inspired by the success of exemplar models in category learning (Kruschke, 1992), associative models (Busemeyer et al., 1997; DeLosh et al., 1997) assume that subjects retain a record in memory with the value of each presented cue during training; when encountering a new cue, subjects simply generalize the criterion value that was associated with similar cues in memory. An immediate problem with associative models is dealing with the phenomena that humans can interpolate to new items with accuracy indistinguishable from original training items (Carroll, 1963; DeLosh et al., 1997; Koh and Meyer, 1991). This consequently led DeLosh et al., 1997 to propose a modification to their original ALM (associative learning model) that could linearly generalize from reference cues in memory when encountering new stimuli (referred to as EXAM[8]). For example, if encountering a new cue between two previously observed cues, EXAM provides a mechanism for interpolating between the line that con-

[8] **EX**trapolating **A**ssociation **M**odel

nects the two stimulus cues in memory[III]. In the case of extrapolation, (to the present author's understanding), EXAM extrapolates along the line connecting the largest or smallest set of training cues in memory[9]. DeLosh et al. (1997) found that the EXAM modification provided relatively excellent fits to human interpolation and extrapolation data (though the modification itself seemed to lack *a priori* theoretical motivation).

Not long after, EXAM's explanatory success was challenged by a counter-intuitive phenomena prevalent in both the FL and CL literature: knowledge partitioning (Lewandowsky and Kirsner, 2000; Yang and Lewandowsky, 2003). Knowledge partitioning is a phenomena where subjects simultaneously generalize conflicting knowledge representations to a stimulus domain, often computationally explained via mixture-of-experts models (Erickson and Kruschke, 1998; Jacobs, Jordan, Nowlan, and Hinton, 1991; Kalish et al., 2004). For example, Kalish et al. (2004) trained subjects on a function structure where most of the training cues followed a positive linear function — with the exception of a few training cues that violated the underlying basis function. Kalish et al. (2004) replicated & extended a very counterintuitive, interesting phenomena[10] where some subjects simultaneously generalized both a positive and negative function during the test phase (see figure 2). Kalish et al. (2004) took their findings as evidence that a single subject can maintain multiple function representations within the same context, and found that their mixture of linear experts model (POLE[11]) was able to account for the phenomenon[12]. Kalish et al. (2004) further demonstrate that DeLosh et al. (1997)'s EXAM model was unable to account for the knowledge partitioning finding (though see McDaniel, Dimperio, Griego, and Busemeyer, 2009 for a follow-up comparative study suggesting potential explanatory weak-

[9] depending on whether extrapolation is in the smaller or larger direction

[10] originally discovered by Lewandowsky and Kirsner, 2000

[11] **P**opulation **O**f **L**inear **E**xperts

[12] Yang and Lewandowsky, 2003 further demonstrate that a mixture of experts model (ATRIUM; Erickson and Kruschke, 1998) can account for knowledge partitioning effects in the context of a category learning experiment.

nesses in the POLE account)^{IV}.

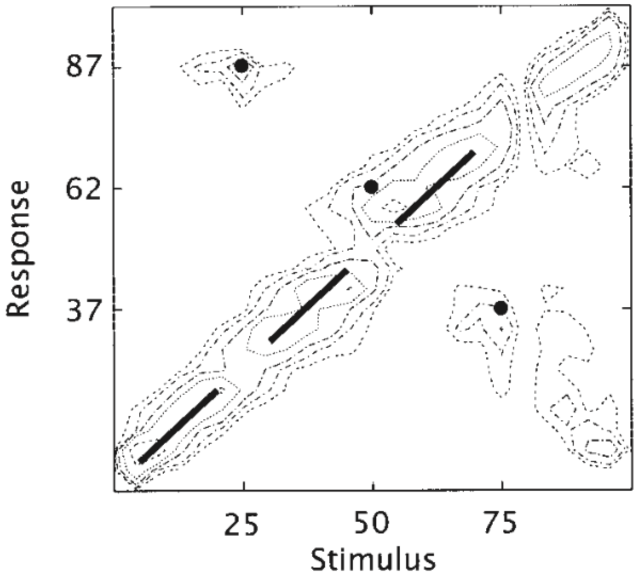

Figure 2: Example of the function structure from Kalish, Lewandowsky, and Kruschke (2004) that elicited knowledge partitioning effects. Taken directly from Kalish, Lewandowsky, and Kruschke (2004).

The debate between associative and 'rule-based' models appears unresolved[13]. While EXAM's success can be taken as support for associative models, it explicitly relies on a representation of linear functions when making interpolation and extrapolation decisions — which seems like evidence partially supporting 'rule-based' theories. Additionally, while not an empirically backed argument, the requirement of storing each exemplar during an FL experiment seems unnecessarily expensive (computationally), particularly when a single layer connectionist network — i.e., regression — can learn a linear function with 2 parameters (slope and intercept). In addition to existing accounts, Bayesian modeling is another more recent approach that is experiencing some preliminary success in the FL literature (Lucas et al., 2015; Narain, Smeets, Mamassian, Brenner, and van Beers, 2014; C. M.

[13] arguably, much like the debate between rule-based and similarity-based models of categorization

Wu, Schulz, and Gershman, 2020). Given the fundamental importance of function learning in human cognition, theoretical explanations of how humans acquire and generalize functional knowledge will likely be a continuing research challenge for psychologists and computational theorists. Crucial next steps might lie in exploring function learning in the context of other cognitive domains (such as perceptual motor learning; Rosenbaum, Carlson, and Gilmore, 2001), or by learning functional relationships from raw perceptual data (as is currently being done in the categorization literature; Singh, Peterson, Battleday, and Griffiths, 2020).

The next section will provide a similar (albeit briefer) review of the category learning literature, followed by a discussion of potential overlapping phenomena in both CL and FL research.

1.2 Category Learning

Category learning and understanding is a relatively broad field of research. Categories can range from being very simple and easy to describe with rules (e.g., things that are red), naturalistic and difficult to describe with rules (e.g., a bird), or structurally complex and abstract (the concept of evolution[14]). While there are many ways through which category learning can be studied, computational theories of category learning have often relied on a very specific paradigm — referred to as traditional, artificial classification learning (TACL; K. J. Kurtz, 2015).

The structure of a TACL experiment is almost identical to that of a function learning experiment: participants are shown a stimulus — typically one at a time — and asked to make a guess about the category membership of that particular exemplar. Subjects learn

[14] Though, as will be argued in a later section, structurally complex 'categories' might be better thought of as graphs of variables connected by both functional and categorical mappings

via corrective feedback provided after their guess. Similar to a function learning experiment, the training phase is typically followed by a generalization phase, where subjects generalize their categorical knowledge to novel, unseen stimuli[15]. Importantly, the past few decades of research have elucidated how human categorization behavior is sensitive to the particular nuances of an experimental procedure, such as category structure (Ashby and Gott, 1988; Kruschke, 1993; Nosofsky and Kruschke, 1992; Shepard, Hovland, and Jenkins, 1961), curricula of stimulus exposure (Kornell and Bjork, 2008; K. H. Kurtz and Hovland, 1953), learning objective (Chin-Parker and Ross, 2004; Kattner, Cox, and Green, 2016), and instructions & context (K. J. Kurtz, Levering, Stanton, Romero, and Morris, 2013). While a TACL experiment is a somewhat limited model of what real-world category learning is actually like, it has provided a number of interesting insights into the constraints learners face when trying to map observable stimulus features onto discrete labels.

Like the FL literature, the CL literature relies on a number of core phenomena that serve as benchmarks for computational explanations of categorization. One common benchmark comes from Shepard et al. (1961), who trained subjects to map 3 binary-valued predictors (i.e., features) to a binary category label[16]; specifically, Shepard et al. (1961) explored 6 different kinds of mappings between stimulus values and labels (figure 3, top). Importantly, Shepard et al. (1961) found a well-replicated disparity between the ease in which certain category structures were learned (Nosofsky, Gluck, Palmeri, McKinley, and Glauthier, 1994). One notable finding was that the easiest category structure to learn (by far) was the category structure that only required attention 1 stimulus feature (i.e., unidimensional[V]); see Kruschke (1993) for a similar demonstration[17] (figure 3, bottom). TACL

[15] Typically, category learning studies don't distinguish between interpolation and extrapolation, since exemplars aren't typically sampled from a continuous function. However, the same general idea applies: generalization stimuli can lie inside or outside the range of items observed during training.

[16] 2 possible categories

[17] Ashby and Maddox (2005) describe these two structures (figure 3, bottom) as encompassing 2 different types of category learning problems: rule-based (left) and information-integration (right).

experiments typically leverage stimuli with a small number of manipulatable features (2-5)[18] — which arguably deviates from most naturalistic, perceptual categories. Nevertheless, the literature has generally supported the idea that the challenge of mapping abstract, artificial variables in a laboratory setting gets harder as the number of required predictor cues increases.

[18] This contrasts with function learning studies, which typically only map one cued feature to one criterion feature.

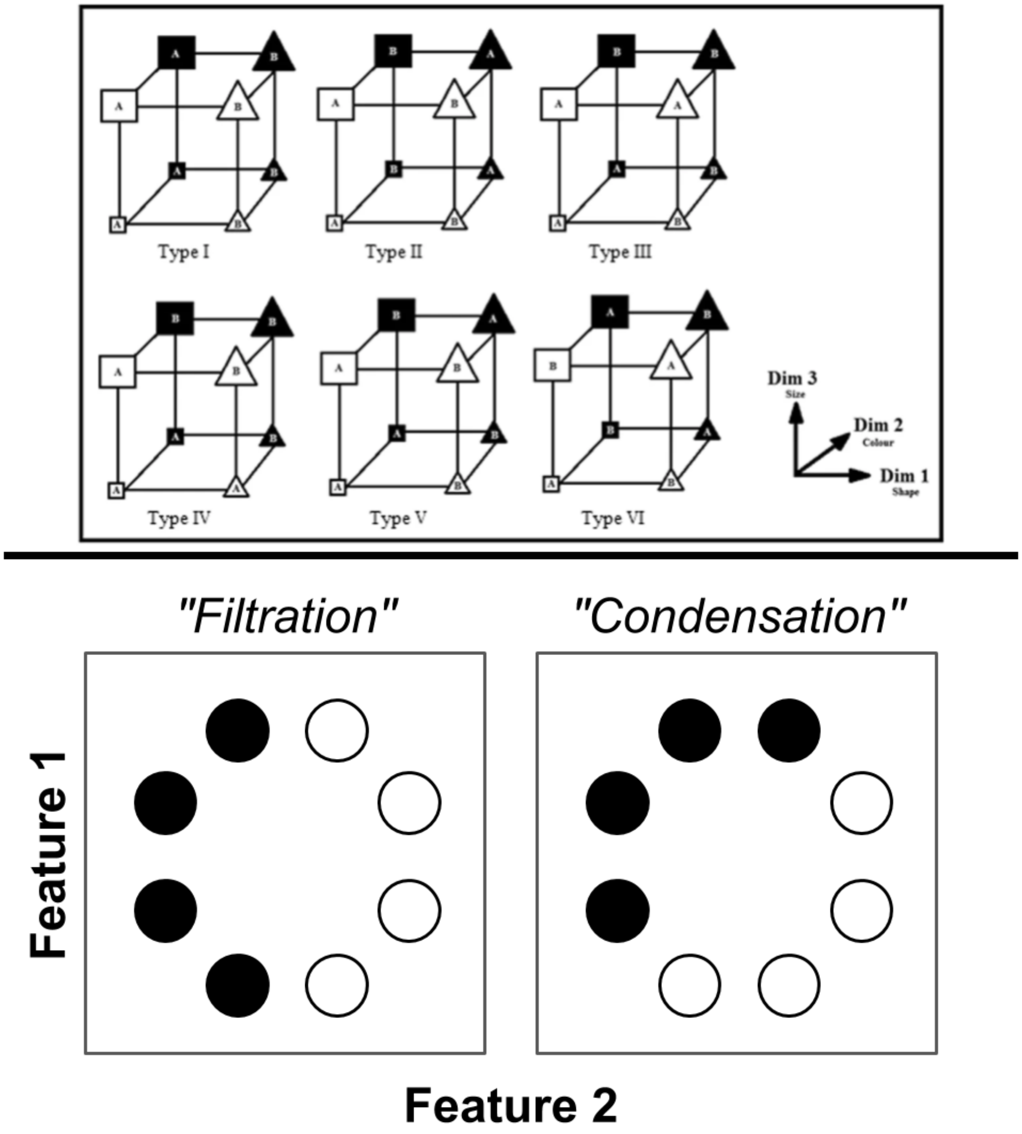

Figure 3: **Top**: Visualization of the 6 different category structures from Shepard, Hovland, and Jenkins (1961) in an arbitrary stimulus domain. Taken directly from Morgan and Johansen (2020). **Bottom**: Visualization of the 'filtration' and 'condensation' category structures used by Kruschke (1993), who found that humans learned the filtration condition (which can be classified on the basis of *a single* feature) easier to learn. Adapted from Kruschke (1993).

Predictions about which categories are easier to learn can also be a liability for some theories. For example, prototype theories predict that multidimensional categories should be easier to learn when they are linearly separable (i.e., can be separated by a multidimensional plane in the stimulus space); in constrast with this prediction, Medin and Schwanenflugel (1981) found that nonlinearly separable categories were as easy for subjects to learn as linearly separable categories[19]. Beyond the issue of ease of learning, another key phenomenon in the CL literature is the Categorical Perception effect (Goldstone, 1994; Livingston, Andrews, and Harnad, 1998; Pothos and Reppa, 2014). In one demonstration, Goldstone (1994) queried subjects ability to perceptually discriminate between various stimulus pairs utilized in a TACL experiment. Critically, Goldstone (1994) found that subjects were much better at distinguished stimuli (post-learning) when those stimuli varied along diagnostic category boundaries (see Goldstone, Lippa, and Shiffrin, 2001 for a replication with similarity ratings)[20].

This section by no means covers the full variety of phenomenological effects uncovered by the TACL paradigm (though more will be discussed in a later section). Further, the TACL paradigm itself only encompasses a small scope of findings from category and concept learning literature as a whole (see Murphy, 2004 for a more thorough review). The next section will provide a brief (and again, incomplete) review of classic models of category learning (with a focus on those that inspired computational modeling in the FL literature).

[19] using stimuli defined by binary features

[20] Categorical perception effects provide an interesting benchmark for computational models, given that the phenomenon was popularized in the CL literature *after* the formulation of many leading classic theories; meaning, any theoretical explanation of category perception effects would have to result from an *a priori* prediction. In other words: classic theorists couldn't cheat).

1.2.1 Models of Category Learning

In the FL literature, models are typically described as falling into two camps: rule-based (curve fitting & regression) and associative (stimulus generalization). The CL literature contains a similar division between models that leverage either logical rules (Levine, 1975; Nosofsky, Palmeri, and McKinley, 1994), similarity to data-driven category representations (K. J. Kurtz, 2007; Love, Medin, and Gureckis, 2004; Minda and Smith, 2001; Nosofsky, 1986), or both (Erickson and Kruschke, 1998). Like models of FL, connectionism has been a fundamental framework for mediating different computational theories of categorization (Erickson and Kruschke, 1998; Gluck and Bower, 1988; Kruschke, 1992; K. J. Kurtz, 2007). Associative and data-driven models of FL and CL also share the property of being implemented as a feed-forward, directed mapping between stimulus cues and responses.

Bayesian cognitive modeling is another, implementation-agnostic, cross-domain framework that is gaining popularity in both the CL and FL literatures (Anderson, 1991; Griffiths, Sanborn, Canini, and Navarro, 2008; Sanborn, Griffiths, and Navarro, 2006). The Bayesian, or 'rational' approach to categorization assumes that learners' objective during a CL scenario is to maximize the probability of successfully inferring the properties of novel objects using whatever representational mechanisms are available (Griffiths et al., 2008). While classic theorists spent a great deal of effort debating the true nature of category representations in the human mind — such as exemplar-based representations (Nosofsky, 1986), prototype-based representations (Minda and Smith, 2001), or something in between (Love et al., 2004; Rosseel, 2002; Vanpaemel, Storms, and Ons, 2005) —, Bayesian theorists assume learners are endowed with the mechanism to produce any arbitrary representational model that optimally captures the available data. Importantly, Bayesian theories don't necessarily conflict with existing frameworks, and can be fluidly integrated into other, more process-centric accounts (like connectionist models; Neal, 2012).

2 Key Areas of Present and Potential Overlap

Computational explanations in both the CL and FL literatures share key elemental principles: (a) both rely on error-driven, trial-wise learning, (b) both constrain learning environment to a small number of easy-to-identify latent variables (serving as cues, features, response, and category labels), and (c) both rely on the same theoretical frameworks for designing computational models. The obvious difference is that the mappings between predictor and response variables are continuous in FL and discrete in CL. Another methodological difference is that FL studies often involve a single predictor variable, where as the stimuli in CL experiments contain multiple predictive features. Further, the predictor and response variables in FL studies are usually described as *attributes* or *properties* (such as weight, size, test scores, etc.); in contrast, the response variables in CL studies are typically framed as conceptual entities[21]. One perspective is that the differences between CL and FL studies are largely in regards to methodological nuances, and both might evoke the same cognitive processes. This section will explore this perspective further to try to address the question of whether CL and FL rely on common mechanistic principles.

2.1 Prior Knowledge

One possible phenomenological overlap between CL and FL is the impact of prior knowledge. In the CL literature, a variety of studies have demonstrated the impact of prior knowledge during category learning experiments. One demonstration comes from Wattenmaker, Dewey, Murphy, and Medin (1986), who found that subjects were better at learning linearly separable (relative to nonlinearly separable) categories only when stim-

[21] This aspect alone could invoke different forms of prior knowledge; in other words, learners might enter a CL or FL study with the expectation that feature-to-feature mappings in the environment are typically described by continuous functions, while feature-to-entity mappings are typically described by discrete probability density functions (Ashby and Alfonso-Reese, 1995)

ulus features shared a common underlying theme (e.g., all related to the feature of "being honest"). The opposite effect emerged when features where framed as being independent[22]. In another demonstration, Heit (1994) found that subjects were better at inferring relations between discretely-valued features when the feature relationships were congruent with prior knowledge[23]. In a very similar preparation, Heit (1995) found that that same effect of prior knowledge slowly diminished as learning progressed; that is, subjects eventually updated their knowledge to match the unintuitive feature relationship defined in the experiment (similar to a finding in the FL literature from work by Sniezek, 1986).

While not particularly surprising, subjects in both FL and CL experiments appear to leverage prior biases about how functional and categorical variables interact[24]. Further, subjects are eventually able to update their prior beliefs in response to experimental feedback. This common phenomenon doesn't necessarily suggest a mechanistic overlap between CL and FL in humans; however, it does have implications for CL and FL theories. Speculatively, the representations underlying human CL and FL may be driven by the nature of prior experience in everyday environments rather than the nature of cognitive mechanisms entirely[25]. However, to the author's knowledge, there doesn't appear to be any systematic investigation of the nature of functional and categorical relationships between psychologically-relevant variables in ecologically representative environments[VI].

2.2 Inference Learning

In a standard TACL experiment, subjects are given a set of stimulus features, and asked to decide which category it belongs to. An alternative learning objective — known

[22] Wattenmaker et al. (1986) used abstract descriptions of stimuli rather than stimuli defined by perceptual characteristics, which may be pertinent to the generalizability of their findings.

[23] For example, pairing the feature *shy* with the feature *avoiding parties* would be congruent with prior knowledge, while pairing the feature *shy* with *attends parties* would be at odds with prior knowledge.

[24] In fact, this particular bias of prior knowledge is often explicitly guarded against when designing experiments (K. J. Kurtz, 2015).

[25] Which partially aligns with arguments made by Bayesian theorists (Anderson, 1991; Griffiths et al., 2008)

as *inference learning* (Chin-Parker and Ross, 2004; Yamauchi and Markman, 1998) — directs subjects to try and guess the value of a stimulus feature *given* a presented category label (as well as the rest of the stimulus features). Rather than mapping stimulus features to category labels, subjects are guided to focus on the within-category statistical relationships between features (see K. J. Kurtz, 2007 for an example of a connectionist model that learns in a similar way). Chin-Parker and Ross, 2004 found that inference learning had important implications on what subjects learned stimulus exposure: inference learners were more sensitive[26] to *prototypical features*, whereas classification learners were more sensitive to *diagnostic features*[27] (see Yamauchi and Markman, 1998 for an earlier demonstration of this phenomenon). Inference learning is a particularly interesting learning objective given that it is essentially the same as a function learning objective, except (1) there is the presence of a discrete cue (the category label), (2) feature values are typically binary instead of continuous (though there's no reason why they need to be[28]), and (3) a single feature often alternates between being a *predictor* or a *predicted* cue[29].

The key finding from the inference learning literature is that the nature of category representations are influenced by the objective subjects are trying to solve (Chin-Parker and Ross, 2004; Ell et al., 2020; Yamauchi and Markman, 1998). The inference learning literature also demonstrates how category learning and function learning can be fluidly integrated into the same experimental preparation. Further, subjects ability to seamlessly integrate both *feature→category* and *feature→feature* mappings in the same context highlights an innate predisposition to do so. One (more cautious) interpretation might con-

[26] Sensitivity to stimulus features was measured via a classification task using stimuli with partially occluded features.

[27] Prototypical features are features shared by the majority of category exemplars; diagnostic features are features that help differentiate category labels. Both are orthogonal but neither are mutually exclusive.

[28] Ell, Smith, Deng, and Hélie (2020) seem to be the only example of an inference learning objective on continuously valued features in the CL literature.

[29] Relatedly, Surber (1987) found that subjects trained to map feature 1 → 2 *do not* consequently learn the inverse mapping between feature 2 → feature 1

clude that the mechanisms that underlie category and function learning in humans are designed to interact. Another (more extreme) interpretation is that the underlying mechanism of CL and FL are identical — the only key difference being *distributional properties of* and *statistical relationships between* latent variables as they typically occur in a learner's naturalistic environment. This extreme interpretation will be explored further in a later section (3.1).

2.3 Dimensionality

Shepard et al. (1961)'s and Kruschke (1993)'s demonstrations highlight the phenomenon where learning is more difficult as the number of relevant dimensions increase. This might seem unsurprising, but it highlights an explanatory limitation for traditional neural network models of cognition (which don't seem sensitive to sparsity of predictive features); Kruschke, 1993). Kruschke (1993) took this as evidence of a selective attention mechanism in human categorization. While there is no direct analog in the FL literature (see 2.4 for a proposal), there is a number of empirical demonstrations that show humans are sensitive to dimensionality (or, number of predictive cues) in a function learning experiment. One such example is the *cue competition effect*, where subjects have difficulty integrating the combined predictability of 2 cues when one cue is more predictive than the other (Busemeyer, Myung, and McDaniel, 1993; see Kruschke and Johansen, 1999 for a review of cue competition effects in the probabilistic category learning literature).

The cue competition effect seems counterintuitive, given that the statistically optimal thing to do in a *multiple cue probability learning* experiment is to integrate as many predictive features as possible. However, humans in both CL and FL preparations seem to selectively weight various features, even when those features are statistically correlated (Gluck and Bower, 1988). One explanation might be that the mechanisms for forming functional and categorical mappings between latent variables are cognitively demanding, and that there is an implicit bias to reduce the dimensionality of the learning domain — thereby freeing up computational resources. A more specific explanation is provided by Kruschke and Johansen (1999), who argued that attentional biases in cue learning are optimal for forming useful representations from a relatively limited number of case exposures. Another alternative (but not mutually exclusive) explanation — which will be discussed in more detail later (3.2) — is that CL and FL in humans is subsumed by a larger computational challenge of global knowledge organization. Specifically, if the goal of FL and CL is to form predictive mappings between latent variables contained within a large-scale knowledge graph, than the computational efficiency of navigating that graph might benefit from a bias towards sparse connectivity.

2.4 Future Directions for Comparative Investigations

The next section will describe findings from the CL literature that (to the author's knowledge) have yet to be applied to the FL literature, but might help address the question of whether FL and CL share overlapping mechanisms.

Formal Metrics of Function Learning

Computational process models of FL and CL provide mechanistic explanations of why some data are easier / harder to learn than others. An additional branch of theoretical categorization research aims to formally describe why some category structures are more difficult to learn than others — independent of whatever computational processes are trying to learn them (Pape, Kurtz, and Sayama, 2015). For example, Feldman (2000) found that the ease of learning the six category structures from Shepard et al. (1961) was correlated with the boolean complexity of each structure (that is, the simplest boolean circuit that can map a set of binary features to a binary category label). An alternative demonstration comes from Pape et al. (2015), who found that the ordering of learning difficulty was predictive by the information complexity of each structure. While formal metrics don't necessarily provide mechanistic explanations of category learning, they do help elucidate either the *information subjects are sensitive* to or the *learning objective they are trying to solve.*

To the author's knowledge, there aren't any formal metrics of why some functions might be easier to learn than others. However, providing a formal metric to explain function learning difficulty fundamentally contrasts from ongoing efforts in the CL literature — particularly when the stimulus features and category labels are typically assumed to be binary. A formal analysis of function learning might require an analysis of the mathematical or information-theoretic properties of continuous function families, which is far beyond the scope of this paper[30]. However, a potential formal metric could build on some of the assumptions made by existing process models, such as compositionality (Schulz, Tenenbaum, Duvenaud, Speekenbrink, and Gershman, 2017). For example, Kalish et al. (2004)'s POLE model of function learning assumes that subjects represent functional knowledge via a piecemeal of composite, linear functions. If this assumption is accurate, the difficulty of learning certain functions over others might be formalized as the number of linear functions required to adequately minimize predictive error during generalization — in which case, the comparative difficulties in learning various function families (linear, quadratic, exponential, periodic) start to become a bit more intuitive. However, that metric specifically would still leave many questions unanswered, such as the preference of positive linear functions (relative to negative linear functions; Brehmer, 1974; Naylor and Clark, 1968).

Representational Shifts in FL

As discussed earlier, *categorical perception* is the relatively well-replicated phenomena where learner's perceptual descriptions of a stimulus domain shift after exposure to

[30] and this author's ability to do so

a category learning preparation (Goldstone, 1994; Harnad, 1987; Liberman, Harris, Hoffman, and Griffith, 1957; Livingston et al., 1998). For instance, Goldstone (1994) found evidence that TACL-style classification learning improved subjects' ability to discriminate between stimuli that varied along diagnostic boundaries (figure 4a). Further, categorical perception effects are often dichotomized into 2 distinct phenomena: between-category *expansion* (where stimuli from opposing categories are viewed as more dissimilar; easier to distinguish) and within-category *compression* (stimuli from the same category are viewed as more similar; harder to distinguish). The specific directions of representational shifts (like compression and expansion) might reflect the nature of category representations subjects acquire during learning.

While it's hard to confidently infer what subjects are actually perceiving during an experiment[31], the idea that the perceptual system can align itself to the conceptual demands of the learner is a fascinating theoretical concern. To the author's knowledge, categorical perception effects have not been explored using inference learning or function learning preparations[32]. Given that regression-style (or, 'rule-based') theories assume that functional knowledge is represented via basis functions, it seems relatively plausible that perceptual changes (measured via stimulus ratings or discriminability tasks) might occur towards the underlying basis function being learned (figure 4b).

[31] Typically, perceptual changes are measured via stimulus discrimination accuracy or pairwise similarity ratings; whether this tightly corresponds to conscious perception is an open question.

[32] Though if they were, the phenomenon would probably require a different name.

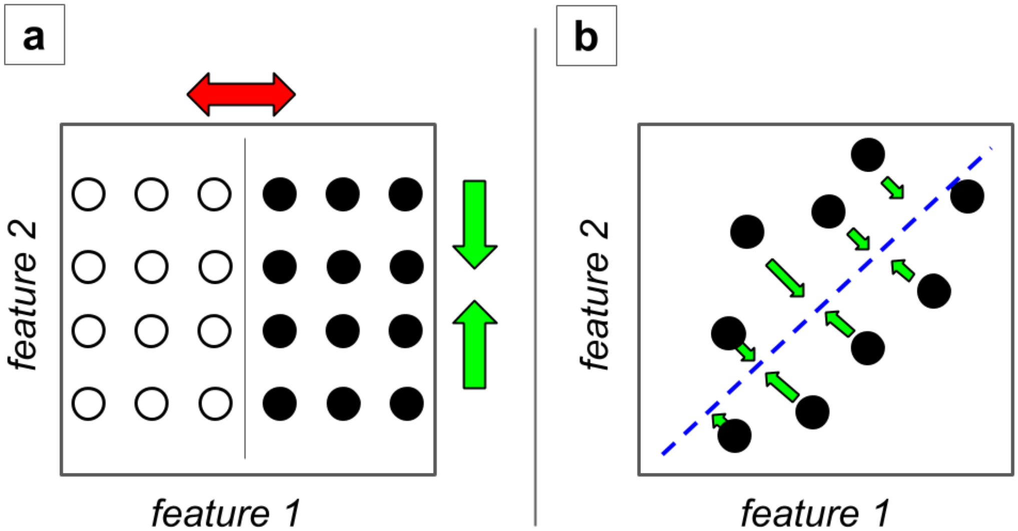

Figure 4: **a**: Example of the category structure used by Goldstone (1994); red arrows indicate *expansion* effects, green arrows indicate **compression** effects. **b**: Visualization of the direction of compression effects one might expect in a function learning task.

FL Replication of Kruschke (1993)

In a relatively impactful demonstration, Kruschke (1993) tested learning of 2 category structures, labeled '*filtration*' and '*condensation*' (figure 3, bottom). Importantly, the stimulus coordinates (or, feature values) in both category structures were identical; the only difference is that in one structure (*filtration*), only 1 stimulus feature was needed to correctly classify each exemplar. In the other structure (*condensation*), each feature was partially predictive of category membership. Kruschke (1993) found that the *filtration* structure was learned faster than the *condensation* structure, which Kruschke (1993) argued was evidence for selective attention in CL. To the author's knowledge, no similar demonstration has been made using an analogous pair of function structures — though such a demonstration might be relatively straightforward to produce. For example, one could train subjects on a 3 dimensional linear manifold; in one case, only 1 feature is needed to accurately predict the criterion value (figure 5a), while in the other case, both features are jointly predictive of the criterion value (5b). If the mechanisms of CL and FL overlap, then one might expect a similar advantage of learning functional patterns that require fewer relevant cues.

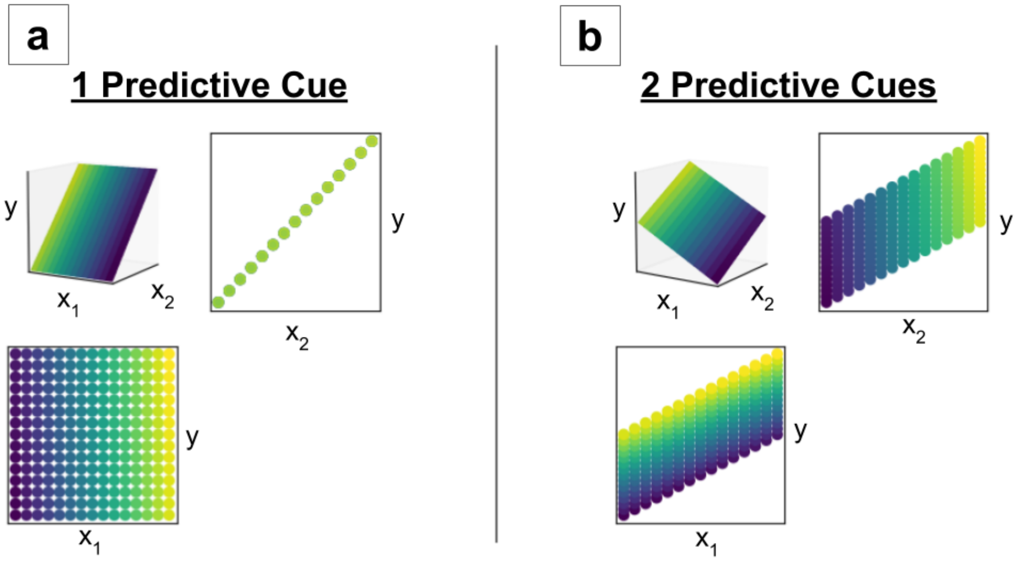

Figure 5: Example of a function learning problem involving 2 continuous predictive cues and 1 continuous criterion. In the first structure (a), only one feature is needed to perfectly predict the criterion (y-axis). In the other (b), both features are needed to predict the criterion. Color indicates a unique identifier for each stimulus coordinate. 3D plots show relationship between all 3 variables; 2D plots show the isolated relationships between each cue X_1, X_2 and the criterion value Y.

3 Further Discussion

3.1 Flexibility of Monotonic Functions

Connectionism (Rumelhart, Hinton, McClelland, et al., 1986) has been a core framework for psychological theories across a variety of domains within psychology, including (but no limited to) categorization (Gluck and Bower, 1988; Kruschke, 1992; K. J. Kurtz, 2007), function learning (DeLosh et al., 1997; Kalish et al., 2004), analogical reasoning (Tomlinson and Love, 2006), and semantic understanding (Rogers and McClelland, 2004). However, the connectionist framework itself is actually relatively ”theory-agnostic”, in that it can be used to instantiate completely divergent theoretical explanations of a psychological construct[33]. Rather, connectionism provides a computational framework for which theories can be instantiated in a brain-style system.

To reiterate, while many leading models of category learning and function learning leverage the connectionist framework (DeLosh et al., 1997; Kalish et al., 2004; Kruschke, 1992; K. J. Kurtz, 2007), they often make very distinct fundamental predictions. As a result, the fact that leading theoretical models rely on connectionism doesn't necessarily provide a cross-domain computational explanation of CL and FL in humans. Interestingly, an activation function commonly used by connectionist models — the sigmoid function (figure 6) — is adequately equipped to handle both categorical and functional mappings (and even capable of approximating the radial basis function used by exemplar models of categorization; Kruschke, 1992; Nosofsky, 1986)[VII]. For categorical decisions, a sigmoid function can take the form of a binary decision boundary indicating the presence or absence of a category member given data from the environment (figure 6a). For function learning, a sigmoid function can approximate linear relationships (figure 6b); in a traditional multilayer neural net, a set of sigmoids can be aggregated to approximate more complex functions[34]. Critically, whether a network of sigmoid functions learns categorical or functional mappings depends on the *data from the environment* it's presented with, bearing resemblance to a differentiable circuit that can approximate both logical operations *and*

[33] In the case of categorization, for example, connectionist models have been used to instantiate models of both exemplar (Kruschke, 1992) and prototype theories (Johansen and Palmeri, 2002).

[34] similar (but distinct) from the approach taken by POLE (Kalish et al., 2004, which piecemeals linear functions per specific datapoints rather than aggregating all functional representations it has access to)

functional associations.

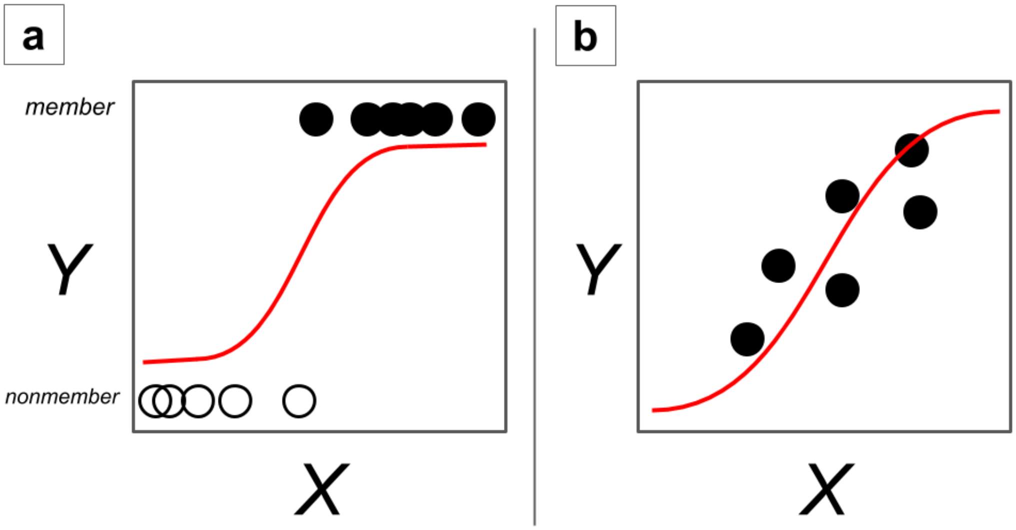

Figure 6: Example of 2 distinct uses of the sigmoid function. **a**: A sigmoid function being used to form a probabilistic decision rule about the presence of a given category given a continuous predictor feature X. Note that the sigmoid function closely approximates the cumulative normal distribution. **b**: The linear portion of a sigmoid function being used to approximate a linear relationship between continuous predictor X and continuous outcome Y. When tuning a sigmoid function via gradient descent, the particular form of the function depends on the distribution of values within the predictor and outcome variables.

While a network of sigmoid neurons is adaptive enough to handle both classification and function learning problems, it's an open question whether this type of model is psychologically meaningful. In the CL literature, multilayer sigmoidal networks have provided poor fits to behavioral data (see Kruschke, 1993 for an analysis about why these failures arise). For example, Kruschke (1993) found that a sigmoidal network (trained via backpropagation) was unable to demonstrate the comparative differences humans exhibit when learning categories diagnosable by 1 or 2 relevant features[35]. However, this empirical failure was alleviated when Kruschke (1993) included attentional weights in the input layer of a standard sigmoidal network (in addition to fixing the weights of the sigmoidal neurons). Further, Kruschke (1993) highlighted that a sigmoidal neural net could approximate exemplar models by using a ”place coding” scheme where input values are discretized and treated as individual features (similar to the approach taken by EXAM; Busemeyer et al., 1997). While there's been little work extending Kruschke (1993)'s sigmoid-based exemplar-approximation model to the rest of the CL literature, it never-the-less highlights the explanatory power of networks of strictly monotonic functions with a magnitude ceiling (i.e., the sigmoid).

But what about function learning phenomena? The empirical failures of rule-based models at predicting extrapolation performance might suggest a preliminary weakness in a regression-style explanation of human function learning. However, previous rule-based models typically utilized relatively complex functions, such as polynomial or exponentials (Brehmer, 1974). An alternative approach — Kalish et al. (2004)'s POLE model — uses a composition of strictly linear functions to explain FL in humans. While POLE can account for knowledge partitioning effects, it fails to account for a key extrapolation phenomenon from McDaniel et al. (2009). To the present author's knowledge, researchers haven't explored whether a composition of sigmoid functions (i.e., a standard, multilayer neural net with sigmoid activations) can explain transfer phenomena in human FL. Interestingly, the extrapolation data from DeLosh et al. (1997) and McDaniel et al. (2009) (experiment 2) are relatively well predicted when fitting the training functions using a combination of 1-2 sigmoid functions (figure 7).

[35] To reiterate, humans learned faster when the categories could be diagnosed on the basis of 1 feature

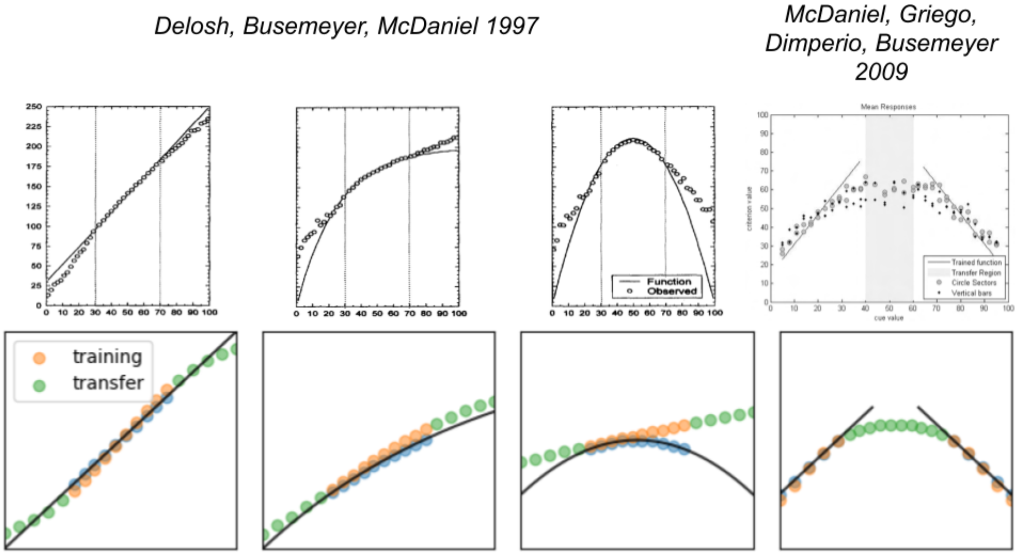

Figure 7: Example of a neural net with sigmoidal hidden units fit to *transfer* data of DeLosh, Busemeyer, and McDaniel (1997) and McDaniel, Dimperio, Griego, and Busemeyer (2009). Top row shows experimental data from humans; bottom row shows neural net predictions. In each function structure, only 1 sigmoid hidden node was used (except for the last function from McDaniel, Dimperio, Griego, and Busemeyer (2009), where 2 sigmoid nodes were fit to both linear training functions). Weights were tuned by the author (not through backpropagation), making this demonstration more of a proof-of-concept than a psychologically meaningful explanation.

A big limitation with the above analysis is that the sigmoid functions were fit post-hoc by the author (making it relatively easy to cheat); what's missing is an a priori system for learning the training data[36]. In addition, these results aren't necessarily surprising given that neural networks are universal function approximators (given enough hidden neurons; Csáji et al., 2001); this alone makes the present explanation unfalsifiable. However, penalizing the number of required sigmoid representations a learner requires might provide a more falsifiable explanation of the nature of extrapolation behavior as well as

[36] The author found that backpropagation alone doesn't seem to be sufficient.

the comparative ease through which some functions are learned over others (by assuming that the difficulty of learning a functional relationship is correlated with the number of sigmoid functions needed to approximate it).

Kruschke (1993)'s demonstration that sigmoidal networks can approximate leading accounts of category learning — in addition to the preliminary, rough demonstration by the present author regarding FL extrapolation with sigmoids — provides some support for the argument that CL and FL overlap mechanistically. Specifically, the cognitive mechanism that underlies learning and generalization of both functional and categorical mappings between latent variables might be explained (in part) as a process of compositional, data-driven approximation using monotonic representations[VIII]. This doesn't, however, explain how humans can acquire verbalizable, algebraic knowledge defining the relationships between latent variables in the environment, which might indicate the necessity for a separate cognitive 'system' (as argued by Ashby, Alfonso-Reese, Waldron, et al., 1998). Alternatively — instead of a separate, distinct system — the feedforward mappings formed through circuits of sigmoid functions might additionally interact with explicit, recurrent memory (which might also leverage sigmoid and *sigmoid-like* functions; Gers and Schmidhuber, 2001)[37]; this might alleviate the need for a dual-systems theoretical explanation[IX].

3.2 Implications of Graphical Knowledge Representation

An important research question regarding the mechanisms of CL and FL in humans (whether those mechanisms are similar or distinct) is how both processes fit within the larger framework of human cognition as a whole. The majority of the categories and functions humans learn are not learned in the laboratory, and the inherent purpose of CL and FL probably isn't to improve accuracy during an experiment. Learning functional and cat-

[37] See also Graves, Wayne, and Danihelka (2014) for an example of an end-to-end, memory-augmented recurrent network that learns to perform explicit symbolic operations.

egorical mappings between variables (which will be referred to as *latent mappings* for the remainder of this section) is clearly useful for navigating through and predicting events in the environment; but, do these learned representations exist as isolated parcels of knowledge?, or are they components within a much larger, more complex representational system? If so, what might that representational system look like?

Graphs are a type of data structure that can be defined as a network of nodes (variables) connected by any type of relationship (Newman, 2003). In addition to being a recently emerging interest in the machine learning literature (Bronstein, Bruna, LeCun, Szlam, and Vandergheynst, 2017; Schlichtkrull et al., 2018; C. M. Wu et al., 2020), graphs have had a rich history in cognitive science. Graphs have also served as the fundamental data structure for early theories of semantic knowledge, dating back to the *spreading activation* theory (Collins and Loftus, 1975; Collins, Quillian, et al., 1969; Quillian, 1967). Graphs (more generally, *network science*) have been and are beginning to be a central part of many psychological literatures, include: language (Vitevitch, 2008) & statistical learning (Lynn and Bassett, 2019; Saffran, Newport, Aslin, et al., 1996), procedural motor learning (Kahn, Karuza, Vettel, and Bassett, 2018), semantic memory (Abbott, Austerweil, and Griffiths, 2012), causal cognition (Danks, 2014), and analogical reasoning (taking the form of propositional networks; Gentner, 1983) — to name a few. When thinking about how CL and FL representations fit into the broader scheme of cognition, it seems that graphical representations are a worthwhile (but speculative) assumption to start with.

What implications would graphical representations have for theories of CL and FL[38]? First, we can assume that the cues, category labels, and stimulus features used in a standard CL or FL procedure all invoke typical latent variables subjects' find meaningful and disparate (otherwise, they probably wouldn't work in an experiment). Then, we can hy-

[38] beyond Danks, 2007 early, innovative demonstration that may theoretical models of category learning could be described as directed graphical models

pothesize that a supervised training procedure in both FL and CL involves learning a probabilistic or rule-based mapping between those latent variables[39]; whether the mappings learned are functional or categorical depends on (at least in part) the latent variables' distributions of values as well as the biases invoked during the context of the learning task. For example, during a category learning experiment, a subject might be told that a set of stimulus features '*is a*' member of a given category (or, *'belongs to' category* X)[40]. This language might immediately cue learners as to what type of latent mapping they should expect to observe. Individual differences in CL and FL studies might also be influenced by the types of mappings each subject has a history of observing (in addition to other variables like fatigue, stimulus presentation order, etc), or what latent variables get activated by the experiment context.

The literature discussed thus par in this paper has focused on a specific type of FL and CL learning procedure — *supervised learning* — where subjects are explicitly told the correct values mapping cues / features to criterion variables / category labels. The premise of this section is to say that these operations (both supervised CL and FL) correspond to a specific type of graphical operation: *edge learning*[41]. However, supervised learning is not the only relevant process in CL and Fl. For instance, in an unsupervised category learning experiment^{X} (Pothos et al., 2011), subjects are not explicitly told which category an item belongs to. Rather, subjects are presented with stimuli and tasked with determining the category labels on their own (typically by sorting a set of stimuli into groups they find most intuitive; Medin, Wattenmaker, and Hampson, 1987; Pothos et al., 2011). This type of learning experience seems quite distinct from supervised learning of

[39] Importantly, it doesn't necessarily matter whether these mappings are learned via multiple systems (Ashby et al., 1998), multiple representations (Erickson and Kruschke, 1998; Kalish et al., 2004), or via data-sensitive applications of a single type of monotonic function (as discussed in the previous section); making this conceptual framing somewhat theory-agnostic at the level of isolated mappings.

[40] In the case of function learning, the subjects might hear language like '*How much?*' or '*How Many?*'

[41] where predictor and predicted variables are nodes, and the relations between them are edges

latent variable mappings, and might be better described as a process of *latent variable discovery*; that is, the generation of new nodes within a larger knowledge graph. The structural context that describes a subjects' history with latent variables in the environment might be a key explanatory variable in why certain category structures are intuitively preferred over others during unsupervised acquisition.

Beyond simply re-framing the problem, what theoretical advantages are gained by proposing graphical representations as a central data structure for both CL and FL? The paradigm through which CL and FL are investigated in the lab attempts to reduce the learning context down to a few isolated variables with the goal of invoking as little prior knowledge as possible (K. J. Kurtz, 2015). However, if the goal of CL and FL research is explanatory generalizability, then it's important to focus on how the phenomena and principles that describe experimental data apply to real world knowledge. Given that many real world knowledge domains seem to embody network structure — e.g., semantic knowledge (Collins, Quillian, et al., 1969; Steyvers and Tenenbaum, 2005), phoneme structure in the English language (Vitevitch, 2008), causal knowledge (Danks, 2014) — then thinking about CL and FL as computational operations on a graph might help explain how people learn categorical and functional relationships in real-world domains. Additionally, graphical data structures provide valuable information for learning and inference at the *level of individual variables*. For example, if my goal was to infer the attributes of *person X*'s social media profile, then I might benefit from looking at the features and attributes of the people that *person X* is connected to. The machine learning literature is currently demonstrating that graphical data structures are very useful for inference and classification of category exemplars (which are arguably the objectives of theoretical models of CL and FL); it therefor seems reasonable (but again, speculative) that the human mind might leverage graphical data structures as well.

To summarize, this section posits the idea that supervised CL and FL in humans both correspond to computational operations between latent variables in a larger graphical representation of knowledge. Unsupervised learning can then be thought of as *graph construction* (Z. Wu et al., 2020), where new nodes are both added and mapped to existing nodes (or, cues / features). Making this theoretical assumption provides a medium through which CL and FL can be framed in the larger context of other cognitive literatures, and explains how humans can make inferences about their environment using information *beyond* strictly the variables they are presented with (leveraging the graphical structure of the domain that is invoked). Graphical representations also provide a shared representational structure that might explain how knowledge is transferred across domains — an assumption already leveraged heavily by theoretical models of analogical reasoning, which typically invoke *propositional* graphs as a fundamental representational structure (Gentner, 1983; Hummel and Holyoak, 1996). An exciting new step for cognitive modeling might lie in recent advances in applying data driven approaches (like deep learning) to graph structured learning problems using the same connectionist principles that constrain many psychological models (Schlichtkrull et al., 2018; Z. Wu et al., 2020).

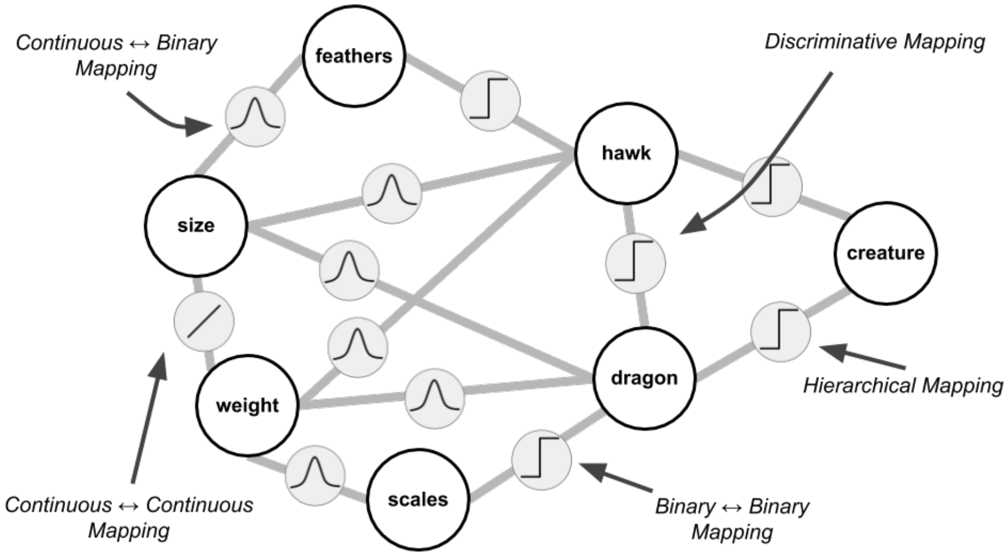

Figure 8: Example of a theoretical knowledge graph, where cues, features, and category labels are all treated as latent variables in a graph. The type of mapping between variables would be determined on each variable's distribution of values (as well contextual factors, experiment instructions, individual biases, etc). Note that these different types of mappings can all be approximated with a sigmoid function.

3.3 Relationship to Other Branches of Cognitive Science

It's beyond the scope (and ideal length) of this paper to provide an exhaustive review of how CL and FL in humans relates to other key branches of the psychological literature; this section will simply highlight a few attempts that have already been made. For instance, there have been a number of previous demonstrations highlighting the link between theories of categorization and theories of associative conditioning in animals (see Gluck and Bower, 1988; Kruschke, 2001). The function learning literature does not seem to have been the subject of such an endeavor; however, it's close ties with the multiple cue probability learning literature suggests a link as well (Busemeyer et al., 1997; Kruschke and Johansen, 1999). There has also been suggestions of an overlap between phenomena in

FL literature and the procedural motor learning literature Rosenbaum et al., 2001). This link is a particularly interesting suggestion, given that it provides support for a speculative hypothesis that abstract intelligence might have evolved from — or co-evolved with — the original mechanisms for spatial navigation in complex physical terrains. In other words, the cognitive mechanism that allows a person to apply the correct amount pressure to the gas petal of a car[42] might be the same underlying mechanisms that allows a person to understand the relationship between energy use and monetary expense (the fundamental difference being whether the latent variables are perceptual or symbolic).

3.4 Conclusion

The following paper aimed to explore the question of whether function learning and category learning in humans leverage the same fundamental cognitive mechanisms. While CL and FL diverge in important ways, an underlying premise of this paper is that those divergences are not necessarily driven by core differences in underlying processing, but rather by the nature of variables in the environment and categorical & functional relations between them.

Overapplying Occam's Razor to psychological theories might be particular problematic given that human behavior is very complex and exists within an environment with a diverse range of evolutionary demands. This makes the basic premise of this paper a bit of a risky position to take. That being said, recent deep learning applications have demonstrated that a relatively simple, universal associative learning mechanism on a network architecture is flexible enough to perform a wide range of behaviors with rather impressive proficiency, including: objective classification (He, Zhang, Ren, and Sun, 2015), object localization (Bazzani, Bergamo, Anguelov, and Torresani, 2016), language interpretation and production (Jawahar, Sagot, and Seddah, 2019; Wang, Chen, and Lee, 2018), and even

[42] This example of a function learning task in the real-world was originally used by Lucas et al., 2015.

reinforcement learning and conditioning (Mnih et al., 2013). This is by no means an argument that modern deep learning is the fundamental explanation of human cognition; rather, it's a proof-of-concept that a relatively simple mechanism for learning and representation can perform impressive feats of cognitive intelligence[XI]. And, while it's relatively appealing to assume that many aspects of cognition are different instantiations of the same underlying mechanisms, it is an open question whether evolution felt the same way.

Notes

[I] This paragraph compares the ease of learning between linear, nonlinear, monotonic, non-monotonic, cyclic, and noncyclic function families. However, it's important to note that these are large, flexible families of functions, and the experimental investigations used to generalize the claims in this section only sampled a few functions from each family. In addition, some functions can be in multiple families. While the empirical results discussed are well-replicated, there may be nuances in concluding which function families are easier to learn than others. What would be useful is a more general metric of function complexity/difficulty that doesn't explicitly rely on traditionally-defined function categories. In addition, the psychophysical scales of the features used in function learning also raises concerns about whether the generalizability of existing findings are limited to the original scales that were utilized (Brehmer, 1974).

[II] An interesting alternative to early, regression-based models might be the autoencoder, which utilizes multiple layers of simple, linear-like functions to produce any particular nonlinear response (importantly, optimized via error-driven learning in a manner consistent with the typical training experience of human learners).

[III] ALM and EXAM are implemented as associative, connectionist networks. One unusual property they share (relative to traditional connectionist networks) is the use of separate neurons for each discretized value of the stimulus dimension. For example, if coding for size ranging from 1mm to 10mm, ALM/EXAM might utilize 10 input neurons for each incrementing value. In more typical neural network (Gluck and Bower, 1988; Kruschke, 1992; K. J. Kurtz, 2007), only a single neuron is required to represent an entire stimulus scale. It's an open question whether ALM/EXAM's representational format is needed to accomplish the same theoretical predictions.

[IV] Notably, mixture-of-experts models are somewhat agnostic to what the actual 'experts' actually are. The knowledge partitioning phenomena doesn't necessarily contradict existing frameworks on it's own; rather, it can be taken as a theoretical argument about the role that contextual gating should play in computational models of category and function learning (similar to how dimensional attention has become an almost ubiquitous property of models of category learning; Erickson and Kruschke, 1998; Kruschke, 1992; Love et al., 2004; Waldron and Ashby, 2001).

[V] Some researchers have conjectured that an ideal category structure should efficiently maximize within-category similarity and minimize between-category dissimilarity (Medin et al., 1987); the unidimensional preferences seems to be at odds with that conjecture.

[VI] Though Rosch and Mervis (1975) argue that *family resemblance* is a particularly common property of naturalistic categories; interestingly, however, *family resemblance* structures seem to lack any particular bias in artificial, low-dimensional stimulus domains (Medin et al., 1987; Shepard et al., 1961).

[VII] The sigmoid function might also be particularly useful because it can approximate an analog→digital conversion (or vice versa), which might shift the computational complexity of certain problems (an idea discussed earlier by Harnad, 1987).

[VIII] Why might monotonic functions with a magnitude ceiling be useful for network-style learning? One possible explanation is in regard to mechanism through which representations are updated during learning. Assuming the mind is a network doing anything akin to backpropagation (see Lillicrap, Santoro, Marris, Akerman, and Hinton, 2020), then the plasticity of neurons are influenced by the 'loss landscape' (Li, Xu, Taylor, Studer, and Goldstein, 2018) defined by the network's optimization goal. Monotonic functions might be useful for producing smooth landscapes (with less local minima) that are easy for the network to traverse (relative to non-monotonic radial basis or periodic functions; Parascandolo, Huttunen, and Virtanen, 2016).

[IX] See also Tsukimoto (2000)'s method for approximating boolean functions from sigmoidal neural nets.

[X] To the author's knowledge, there is no analogous, unsupervised preparation in the function learning literature. It's unclear what impact such a preparation would have on subjects' ability to learn functional relationships. Speculatively, the presence of functional relationships might predict what latent variables subjects find most important. For example, given an environment with a large set of potential variables of interest, variables that have an obvious relationship (either categorical or functional) might be especially relevant in the future — in that they allow subjects to make predictions in situations where information is limited.

[XI] despite being still far behind the capabilities of human adults (Lake, Ullman, Tenenbaum, and Gershman, 2017)

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. In *Nips* (pp. 3050–3058).

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, *98*(3), 409.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, *39*(2), 216–233.

Ashby, F. G., Alfonso-Reese, L. A., Waldron, E. M., et al. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, *105*(3), 442.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol. 56*, 149–178.

Bazzani, L., Bergamo, A., Anguelov, D., & Torresani, L. (2016). Self-taught object localization with deep networks. In *2016 ieee winter conference on applications of computer vision (wacv)* (pp. 1–9). IEEE.

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 38.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18–42.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks.

Busemeyer, J. R., Myung, I. J., & McDaniel, M. A. (1993). Cue competition effects: Empirical tests of adaptive network learning models. *Psychological Science*, *4*(3), 190–195.

Byun, E. (1996). *Interaction between prior knowledge and type of nonlinear relationship on function learning.* (Doctoral dissertation, ProQuest Information & Learning).

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*(2), i–144.

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 216.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Collins, A. M., Quillian, M. R. et al. (1969). Retrieval time from semantic memory.

Csáji, B. C. et al. (2001). Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, *24*(48), 7.

Danks, D. (2007). Theory unification and graphical models in human categorization. *Causal learning: Psychology, philosophy, and computation*, 173–189.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models.* Mit Press.

DeLosh, E. L. (1995). Hypothesis testing in the learning of functional concepts. *Unpublished master's thesis, Purdue University, West Lafayette, IN.*

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968.

Ebersbach, M., Van Dooren, W., Van den Noortgate, W., & Resing, W. C. (2008). Understanding linear and exponential growth: Searching for the roots in 6-to 9-year-olds. *Cognitive Development*, *23*(2), 237–257.

Ebersbach, M., & Wilkening, F. (2007). Children's intuitive mathematics: The development of knowledge about nonlinear growth. *Child development*, *78*(1), 296–308.

Ell, S. W., Smith, D. B., Deng, R., & Hélie, S. (2020). Learning and generalization of within-category representations in a rule-based category structure. *Attention, Perception, & Psychophysics*, 1–15.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155–170.

Gers, F. A., & Schmidhuber, E. (2001). Lstm recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, *12*(6), 1333–1340.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*(2), 125–157.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43.

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric bayesian density estimation. *The probabilistic mind: Prospects for Bayesian cognitive science*, 303–328.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In *Categorical perception: The groundwork of cognition* (pp. 1–52). Cambridge University Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the ieee international conference on computer vision* (pp. 1026–1034).

Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1264.

Heit, E. (1995). Belief revision in models of category learning. In *Proceedings of the seventeenth annual conference of the cognitive science society* (pp. 176–181). Erlbaum Amsterdam.

Hummel, J. E., & Holyoak, K. J. (1996). Lisa: A computational model of analogical inference and schema induction. In *Proceedings of the eighteenth annual conference of the cognitive science society* (pp. 352–357).

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, *3*(1), 79–87.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does bert learn about the structure of language?

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive psychology*, *45*(4), 482–553.

Kahn, A. E., Karuza, E. A., Vettel, J. M., & Bassett, D. S. (2018). Network constraints on learnability of probabilistic motor sequences. *Nature human behaviour*, *2*(12), 936–947.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological review*, *111*(4), 1072.

Kattner, F., Cox, C. R., & Green, C. S. (2016). Transfer in rule-based category learning depends on the training task. *PloS one*, *11*(10).

Koele, P. (1980). The influence of labeled stimuli on nonlinear multiple-cue probability learning. *Organizational Behavior and Human Performance*, *26*(1), 22–31.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 811.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological science*, *19*(6), 585–592.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*(1), 3–36.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of mathematical psychology*, *45*(6), 812–863.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1083.

Kurtz, K. H., & Hovland, C. I. (1953). The effect of verbalization during observation of stimulus objects upon accuracy of recognition and recall. *Journal of Experimental Psychology*, *45*(3), 157.

Kurtz, K. J. (2007). The divergent autoencoder (diva) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. In *Psychology of learning and motivation* (Vol. 63, pp. 77–114). Elsevier.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of shepard, hovland, and jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*.

Levine, M. (1975). *A cognitive theory of learning: Research on hypothesis testing.* Lawrence Erlbaum.

Lewandowsky, S., & Kirsner, K. (2000). Knowledge partitioning: Context-dependent use of expertise. *Memory & cognition*, *28*(2), 295–305.

Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the loss landscape of neural nets. In *Advances in neural information processing systems* (pp. 6389–6399).

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, *54*(5), 358.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 1–12.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 732.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. *Psychological review*, *111*(2), 309.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic bulletin & review*, *22*(5), 1193–1215.

Lynn, C. W., & Bassett, D. S. (2019). Graph learning: How humans infer and represent networks. *arXiv preprint arXiv:1909.07186*.

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*(2), 668.

McDaniel, M. A., Dimperio, E., Griego, J. A., & Busemeyer, J. R. (2009). Predicting transfer performance: A comparison of competing function learning models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 173.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 355.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive psychology*, *19*(2), 242–279.

Miller, P. M. (1971). Do labels mislead? a multiple cue study, within the framework of brunswik's probabilistic functionalism. *Organizational Behavior and Human Performance*, *6*(4), 480–500.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Morgan, E. L., & Johansen, M. K. (2020). Comparing methods of category learning: Classification versus feature inference. *Memory & Cognition*, 1–21.

Murphy, G. (2004). *The big book of concepts.* MIT press.

Narain, D., Smeets, J. B., Mamassian, P., Brenner, E., & van Beers, R. J. (2014). Structure learning and the occam's razor principle: A new view of human function acquisition. *Frontiers in Computational Neuroscience*, *8*, 121.

Naylor, J. C., & Clark, R. D. (1968). Intuitive inference strategies in interval learning tasks as a function of validity magnitude and sign. *Organizational Behavior and Human Performance*, *3*(4), 378–399.

Neal, R. M. (2012). *Bayesian learning for neural networks.* Springer Science & Business Media.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167–256.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1961). *Memory & cognition*, *22*(3), 352–369.

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. *The psychology of learning and motivation*, *28*, 207–250.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53.

Pape, A. D., Kurtz, K. J., & Sayama, H. (2015). Complexity measures and concept learning. *Journal of Mathematical Psychology*, *64*, 66–75.

Parascandolo, G., Huttunen, H., & Virtanen, T. (2016). Taming the waves: Sine as activation function in deep neural networks.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K. J., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, *121*(1), 83–100.

Pothos, E. M., & Reppa, I. (2014). The fickle nature of similarity change as a result of categorization. *Quarterly Journal of Experimental Psychology*, *67*(12), 2425–2438.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral science*, *12*(5), 410–430.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573–605.

Rosenbaum, D. A., Carlson, R. A., & Gilmore, R. O. (2001). Acquisition of intellectual and perceptual-motor skills. *Annual review of psychology*, *52*(1), 453–470.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*(2), 178–210.

Rumelhart, D. E., Hinton, G. E., McClelland, J. L., et al. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, *1*(45-76), 26.

Saffran, J. R., Newport, E. L., Aslin, R. N., et al. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*(4), 606–621.

Sanborn, A., Griffiths, T., & Navarro, D. (2006). A more rational model of categorization.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. In *European semantic web conference* (pp. 593–607). Springer.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive psychology*, *99*, 44–79.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Singh, P., Peterson, J. C., Battleday, R. M., & Griffiths, T. L. (2020). End-to-end deep prototype and exemplar models for predicting human behavior. In *Proceedings of the annual conference of the cognitive science society.*

Sniezek, J. A. (1986). The role of variable labels in cue probability learning tasks. *Organizational Behavior and Human Decision Processes*, *38*(2), 141–161.

Sniezek, J. A., & Naylor, J. C. (1978). Cue measurement scale and functional hypothesis testing in cue probability learning. *Organizational Behavior and Human Performance*, *22*(3), 366–374.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, *29*(1), 41–78.

Summers, S. A., Summers, R. C., & Karkau, V. T. (1969). Judgments based on different functional relationships between interacting cues and a criterion. *The American Journal of Psychology*, *82*(2), 203–211.

Surber, C. F. (1987). Formal representation of qualitative and quantitative reversible operations. In *Formal methods in developmental psychology* (pp. 115–154). Springer.

Tomlinson, M., & Love, B. C. (2006). Learning abstract relations through analogy to concrete exemplars. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 2269–2274). Lawrence Erlbaum Associates Mahwah, NJ.

Tsukimoto, H. (2000). Extracting rules from trained neural networks. *IEEE Transactions on Neural networks*, *11*(2), 377–389.

Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In *Proceedings of the annual conference of the cognitive science society* (Vol. 27, pp. 2277–2282). Lawrence Erlbaum Associates; Mahwah, NJ.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research.*

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic bulletin & review*, *8*(1), 168–176.

Wang, Y.-S., Chen, Y.-N., & Lee, H.-Y. (2018). Topicgan: Unsupervised text generation from explainable latent topics.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology*, *18*(2), 158–194.

Wu, C. M., Schulz, E., & Gershman, S. J. (2020). Inference and search on graph-structured spaces. *bioRxiv.*

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems.*

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and language*, *39*(1), 124–148.

Yang, L.-X., & Lewandowsky, S. (2003). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 663.