COGS 4800 Michael Woodford

Samarth Agrawal

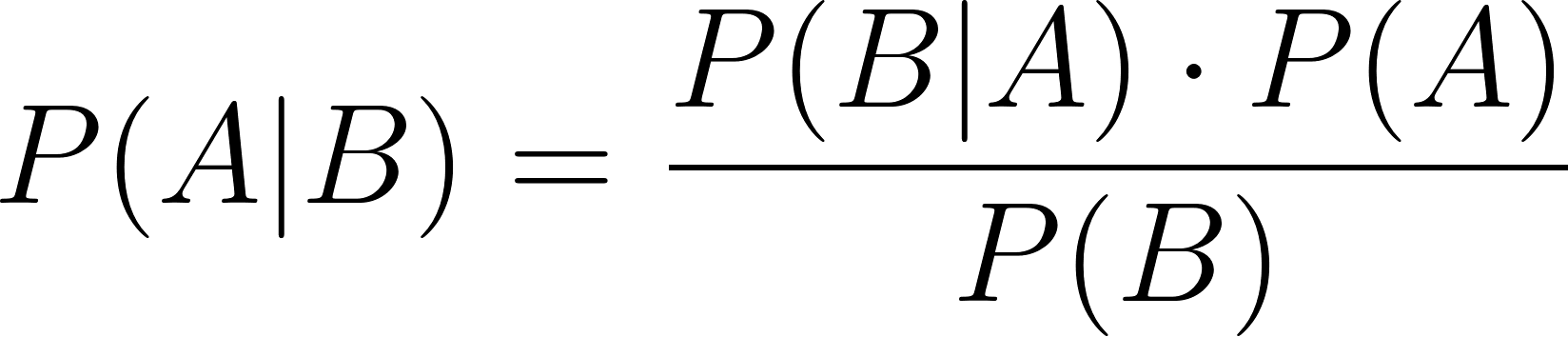
May 12th, 2023

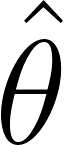
Resource-Rational Identification of Causally Relevant Variables

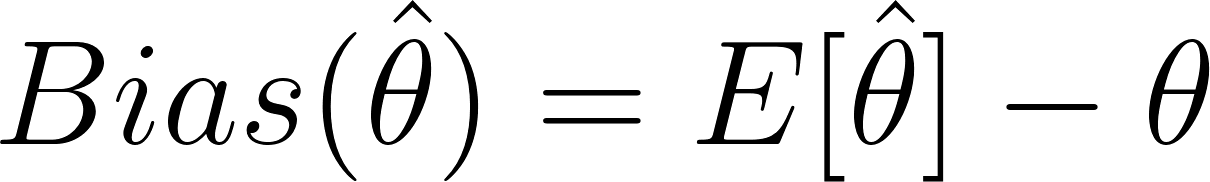
# Introduction

Our lived sensory experience is that of a coherent world with reliable physical laws. In reality, it is more fitting to say the mind *constructs* the world it perceives from fragmented and noisy signals. Before any stimulus enters our conscious awareness, our brain filters and interprets information based on underlying assumptions it has about the relationship between perceptions and reality. Rather than interacting with information in a vacuum, all cognition occurs within the context of the world we expect to see. The term “inductive bias” refers to the set of assumptions that constrain which hypotheses are formed before any data is observed. Humans often rely on inductive reasoning, which derives principles or heuristics from data. There is no logical basis for induction- there are often infinitely many categorizations which all fit a given finite sample of observations. Inductive biases narrow down the hypothesis space and constrain deliberation to those that the cognitive system believes will be predictive of unseen data. The term inductive bias is most prevalent in the field of Artificial Intelligence to categorize how machine learning algorithms generalize and what predictions they will make. For example, all parameterized estimators (such as Linear Classifiers which assume a linear relationship between features and labels) can be viewed as containing strong inductive biases. Such models assume a specific input-output form or function class a priori without seeing any data, which reformulates the task of learning into the task of finding optimal values of pre-specified parameters. For non-parametric models or universal approximators (such as Neural Networks), what conclusions the decision system will reach are less obvious. The inductive biases of less interpretable state-of-the-art architectures remains an active area of research within Artificial Intelligence. [[1]](#footnote-0) [[2]](#footnote-1)

The basic paradigm of supervised learning is quite analogous to that of inductive reasoning and will be the decision setup used throughout this work. Given a limited sample of data points, how well can we abstract an underlying mechanism through which the data might have been generated? Since nomenclature varies across fields, it is helpful to disambiguate various terms that have a similar usage to what is meant by inductive bias in machine learning. In various pscyhology and cognitive science domains centered on probabilistic models, the term overhypothesis refers to “any form of abstract knowledge that sets up a hypothesis space at a less abstract level” [[3]](#footnote-2) and can be viewed as either a specific type of inductive bias or synonymous with the term. From the perspective of Bayesian Inference, priors serve essentially the same purpose as inductive biases. Pre-data expectations are encoded as probability distributions which are then updated by evidence (encoded in a conditional distribution known as the likelihood). The process of Bayesian updates is precisely specified in Bayes Rule, one of the central theorems of probability.

[](https://www.codecogs.com/eqnedit.php?latex=P(A%20%7C%20B)%20%3D%20%5Cdfrac%7BP(B%7CA)%20%5Ccdot%20P(A)%7D%7BP(B)%7D#0)

Kahneman and Tversky’s landmark 1974 paper “Judgment under Uncertainty” investigates subjective estimates of the probability of events in the context of behavioral economics. They catalog “cognitive biases that stem from the reliance on judgmental heuristics” such as the oft-cited representative bias, in which the probability assigned to a person having a specific occupation is dominated by the fit to a relevant personality description. [[4]](#footnote-3) In our language, cognitive bias refers to *statistical* bias. The bias of an estimator parameterized by with statistic [](https://www.codecogs.com/eqnedit.php?latex=%5Chat%7B%5Ctheta%7D#0) is given by

[](https://www.codecogs.com/eqnedit.php?latex=Bias(%5Chat%7B%5Ctheta%7D)%20%3D%20E%5B%5Chat%7B%5Ctheta%7D%5D%20-%20%5Ctheta#0)

This represents the expected discrepancy between one’s estimate and some normatively correct value. Inductive bias in our usage refers to the ‘judgmental heuristics’ themselves which are the source of the predictions that are ultimately made (statistically biased or not). Though priors, inductive biases, heuristics, and overhypotheses operate through context-dependent mechanisms, we view them as different representations of the same underlying concept. Kahneman and Tversky’s work highlights how inductive biases that falsely categorize the ground truth reality of the external world can very easily lead to systematic errors. The prevalence of such harmful cognitive biases might lead one to conclude that an ideal decision maker would not constrain hypothesis spaces in the limiting way that humans default to. However, *inductive biases are indispensable*. Without them, effective abstraction or induction is impossible. [[5]](#footnote-4) Cognitive systems face the Herculean task of trying to make sense of a dazzlingly complex world. Even setting aside computational constraints, it is not even theoretically possible to parse infinite-size hypothesis spaces without some mechanism of disregarding implausible conclusions. The normative question is not *whether* one should discount certain hypotheses before seeing evidence but *how* one should do so while navigating various tradeoffs.

We are most interested in the types of abstract structures and world models that inform which hypotheses are ever considered by humans and how they might inform Artificial Intelligence. How does one appropriately choose the correct inductive bias for a given situation? And what link can be drawn between inductive bias and effective generalization? There is an inherent tradeoff present in how specific one makes an inductive bias. More stringent constraints on considered hypotheses might lead one to learn more efficiently but will apply to a narrower class of situations, potentially leading to the types of errors discussed by Kahneman and Tversky. While there are many different kinds of inductive bias, in this work we examine abstract causal models and how they guide the task of induction. Human beings are wired to view the world through a causal lens, interpreting stimuli in terms of causes and effects. Our existing internal models of how objects in the world interact drive the inferences we make in daily life. Many of the most fundamental and enduring causal inductive biases are formed in early childhood and persist throughout adult life. This means to understand such expectations, it is necessary to dive into child developmental psychology. What kinds of models of the world do babies develop- when and how do they learn these from experience? In this work we are interested in the formation of inductive biases from two perspectives. Firstly, through the lens of child psychological development and how humans learn to make sense of the world. And secondly in a less descriptive and more normative account for the rational construction of explanatory world models and how they may inform computational models of intelligence. Computational accounts for these world models can provide significant insight into the development of artificial intelligence. Understanding which inductive biases a cognitive architecture “should” have can guide Machine Learning practitioners to select models in a more deliberate-design driven way. It can also address increasing concerns about the impacts of Artificial Intelligence on society. For example, OpenAI’s GPT 4 model was trained on 45 gigabytes of data, has over 1 trillion parameters, and consumed energy equivalent to 700 US households to train. [[6]](#footnote-5) The ability to learn with greater sample efficiency and with the stunning computational efficiency of biological cognition could lessen these computational demands, making AI both more sustainable and more democratic. Furthermore, rational constructivist theories of cognition offer the possibility of more interpretable (and thus safer) models whose behavior can more easily be trusted.

The format of the paper is as follows. In Section 1, we provide an Introduction to the primary motivation for this work and introduce the core questions surrounding inductive bias. In Section 2, we explore the Background for the topic, examining the existing literature that has contributed to our understanding of the topic. In Section 3, we motivate and describe a toy illustrative example and empirical study we have conducted that investigates causal inference as an example of a critical inductive bias. Section 4 presents and analyzes the Results of our simulation and possible implications for the questions at hand.

# Background

### Psychological evidence for formation of inductive biases in childhood

We begin by briefly reviewing some experimental results suggestive of inductive biases early in cognitive development. Collectively, a large body of work in psychology and cognitive science provides substantial evidence that hypothesis spaces are narrowed from a very early age, forming a critical component of how humans interpret external stimuli. While the literature on this was not really explored, it is possible that childhood constitutes an “exploration” phase in which notions are formed and readjusted whereas reduced neuroplasticity in adulthood might suggest an “exploitation” phase in which our assumptions are more static. Here we discuss two different studies which themselves draw upon well-established empirical bodies of work.

A study by Eliana Colunga and Linda B Smith, “From the Lexicon to Expectations About Kind” [[7]](#footnote-6) examined the connection between early language development and the formation of hypothetical constructs or concepts instantiated in the form of ontologies. Children are observed to form categories with a homogenous property that in some sense defines the category. A prominent example of this is the shape bias, where physical solids are most readily identified by their shape (due to the hierarchical structure of this assumption, the shape bias is technically more of an overhypothesis than an inductive bias; we will use that term for precision but again assert that these terms are largely interchangeable). When shown an prototypical object and given a novel noun for it, children are then asked to pick the mostly likely second example from a set of options (“this is a dax - which of these are daxes?”). Both children and adults select objects homogenous in shape but heterogeneous in all other attributes – daxes are objects that are dax-shaped, not dax-colored or dax-textured. Whereas for non-solid objects, children infer material to be the homogenous property. Interestingly enough, evidence from a study by Landau et al indicates that these ontological overhypotheses are actually over-applied in children as compared to adults, potentially indicating that early understanding comes from more crude general categories which are revised and imbued with nuance over time. [[8]](#footnote-7) Storing and acting upon abstract categories is far easier than recalling individual object properties. The researchers found that different kinds of prompting affected the strength and nature of the biases the participants had- for example, saying “this is a kind of dax” instead resulted in a weaker shape bias, and allowed for more flexible categorizations and consideration of alternative properties as part of the ontology. Various other forms of prompting caused different overhypotheses to take precedence over the shape bias as well. How the evolution of category and concept understanding ties to evolving linguistic maturity is an active field of research, but not directly relevant.

For our purposes, there are several key takeaways from Colunga and Smith’s study. Firstly, the study indicates that inductive biases do not function in a binary on/off way. Even when faced with the same task, suggestive external insights into how best to interpret the data seem to affect the degree of confidence, and as a result, the strength of the overhypothesis. As mechanisms to constrain the hypothesis space, it is natural that these biases are rooted in a priori expectations about the situation one is facing. The flexibility of how confidently the children were inclined to prioritize shape based on linguistic prompting suggests that inductive biases are responsive to changes in these expectations. In addition to their review of psychological literature, Colunga and Smith ran a series of computational experiments to examine whether associative networks (described in more detail in the “Computational Models” section) were capable of learning overhypotheses concerning solid and non-solid categories based on word-instance pairings. Labeled category names (like “ball”), represented as vector word embeddings, were juxtaposed with information about shape, material, and solidity. Not only were the networks able to learn associations between individual categories and properties, it was able to abstract that solidity implied shape homogeneity and non-solidity implied material homogeneity. Furthermore, the strength of these associations was sensitive to the lexical data generated by the researchers. The ability of Colunga and Smith’s networks to approximate the kinds of generalizations observed in children suggests two things. Not only are inductive biases responsive to expectations, but (encoded as overhypotheses) are themselves flexibly learned from data. The shape bias itself reflects real-world correlations in shape that manifest because of the physical properties of solids. [[9]](#footnote-8) Secondly, the structure of the associative networks (and the parallelism observed in neural computation) suggests a graded web of interconnected hypotheses that are learned in tandem.

While the novel noun generalization task is not explicitly causal, the search for homogenous properties echoes a search for why categories are what they are. If a child hears over and over that a thing is deemed a dax because it is dax-shaped, they are in some sense identifying shape as a causally relevant variable for the object class. This task was popularized by philosopher of science James Woodward as ‘the problem of variable choice’ [[10]](#footnote-9). Woodward notes that before any causal identification or generalization can be done, a system must determine which variables to even consider in the first place. Which phenomena cognitive architectures are wired to interpret as features is inherently tied to existing causal world models. In some sense noting down the wrong features is a non-starter for obtaining any sort of systemic understanding. Out of possibly thousands of different properties that can be deemed salient, can we construct a heuristic for which to pay attention to? Alison Gopnik explores a similar line of questioning in her study “Learning What to Change: Young Children Use “Difference-Making” to Identify Causally Relevant Variables”. [[11]](#footnote-10) As the title suggests, the study found that attention is allocated to those features whose change in value is correlated with a change in the outcome variable of interest. For example, if a room lights up immediately after I flip a switch, I am pretty inclined to infer that the switch was a light switch. But if the light switch turns on the light in a manner that makes the difference making correlation far less obvious (such as after a one hour delay) I am far less likely to attribute electrical causation to the switch. In the study, children were asked to guess which combination of flower-pot pattern and watering can color helped a cactus grow desirably shaped fruit. Children as young as three inferred causal relationships from just three data points, the minimum required to highlight a unique difference making variable. In addition to abstracting causal relationships from very little data, the predictions of the child participants implied that these same causal relationships were extended to novel variables. If pattern A yielded outcome i, then pattern B was assumed to yield an outcome ii, with neither B nor ii having been observed. Goddu et al’s study suggests that causal notions are very easily assumed to hold true in general and form lasting overhypotheses that override any specific associations of particular objects.

### Rational Constructivism

While it may now seem less controversial to suggest that children have beliefs that constrain how they view the world, how this knowledge is acquired and in what form it is represented remains uncertain and subject to debate. In terms of acquisition, theories span a spectrum from nativist accounts that posit certain things to be innate or “hardwired” from birth to empirical accounts which emphasize a “blank slate” which learns from data and experience alone. Substantial debate over where the shape bias lies on the nativist-empiricist spectrum resists resolution, with evidence for both sides. Associative accounts and the ability of contrived experimental setups to alter or weaken the bias suggest empiricism while the relative speed by which the shape bias manifests in infants (even younger than the 3-5 age range) suggests at least some sort of innate predilection. [[12]](#footnote-11)

The cognitive development theory of rational constructivism draws from both of these traditions, asserting that the kinds of knowledge representations and structures that characterize causal models are rather real (a rather detailed overview of the motivations of rational constructivism was published by Alison Gopnik and Henry Wellman; [[13]](#footnote-12) here we highlight the key points). Rational Constructivism offers a theory of learning in which knowledge is not passively absorbed but actively incorporated into abstract yet structured theories of how the world functions. [[14]](#footnote-13) Advocates of the rational constructivist theory seem to have embraced probabilistic models, particularly those representing causal knowledge, as the most convincing accounts for experimental results. [[15]](#footnote-14) Rational constructivism is rooted in a probabilistic model of cognition which allows for innate overhypotheses that specify Bayesian priors at exceedingly abstract levels of knowledge. Such priors might for example support a default readiness to accept shape as a causally relevant property which is still malleable to overwriting. One of the central analogies of this “theory theory” [[16]](#footnote-15) (pg 1086) is that the formation and revision of mental representations of the world follows a very similar process to theory revision in philosophy of science.

### Computational Models of Learned Inductive Biases

Here we review the types of computational models underlying the various theories of inductive bias that we have discussed. Psychological corroboration and the explanatory power of a theory to account for behavior are compelling in their own right. But attaching precise mathematical language to vague theoretical accounts can help isolate which essential features of the theory hint at a greater truth and establish the necessary principles for scientific falsifiability. By their very premise, inductive biases are hierarchical. Beliefs and concepts related to hypothesis spaces are inherently at a higher layer of abstraction than the raw data which hypotheses seek to fit. What models of these abstract cognitive structures can we instantiate in code and do these models provide explanatory power beyond the data that motivated them? Various theories of cognition would appear far more plausible if their relevant mathematical descriptions are able to not just fit existing experimental data, but are predictive of findings that postdate them- if the theory of generalization itself is generalizable. Broadly, there are two approaches to constructing such models- a top down philosophy which seeks to explicitly represent the abstract concepts and schemas that humans seem to be using and a bottom-up “emergentist” approach more focused on how individual units in the structure evolve and update over time.

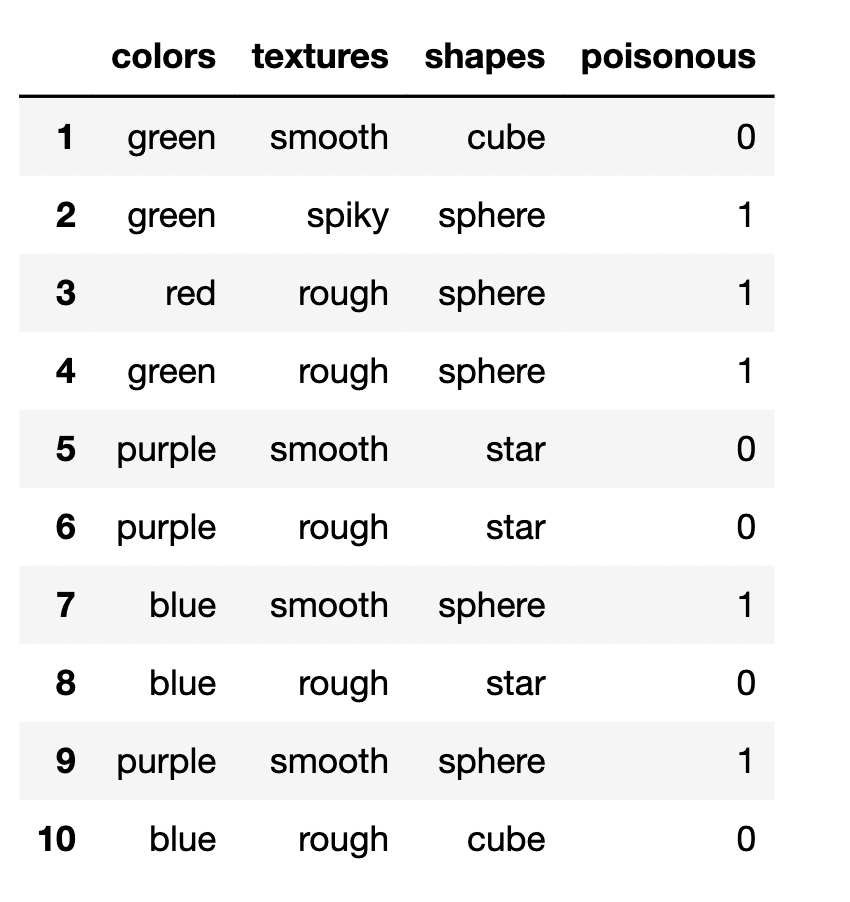
The aptly named emergentist approach conceives of higher level concept manipulation (involving categories, structures, and objects) as an emergent behavior of lower level mechanisms. The ultimate goal of these models remains robustly capturing the underlying relationships between data. However, learning rules are not explicitly represented in any part of the model; abstract higher level concepts (if present in the first place) emerge from the collective behavior of innumerable units. [[17]](#footnote-16) The design of emergentist model architectures tends to be network-oriented and focuses on the mechanistic interactions between nodes. For example, the connectionist neural networks of the deep learning revolution center around activation functions that dictate how neurons “fire” and the backpropagation algorithm which describes how neuron weights (a rough analogue for synaptic connections) update. Associative learning (used in the aforementioned work on novel noun categories) falls more directly under this emergentist umbrella and is even more biologically inspired. Colunga and Smith used associative models known as Contrastive Hebbian Networks of varying sophistication. Hebbian learning, named after neuropsychologist Donald Hebb, refers to a class of algorithms that mimic how neuron connections in the brain grow stronger as they are repeatedly exposed to similar stimuli. A Hebbian network learns by strengthening connection weights between units activated at the same time. Over time, the strongest connections are reflective of genuine statistical relationships between variables (hence the name “associative” learning). A close cousin of Hebbian networks known as Hopfield networks are wired for pattern recognition and are thus extremely useful for studying the formation of memories. The associative models used in the Colunga and Smith model indeed utilized Hebbian Learning but were not strictly emergentist due to the explicit modularity of the graphs, partitioning some nodes as representing shape, others designating solidity, material, and other abstract properties more typical of top-down models.

Rational constructivists tend to favor top down probabilistic models, particularly those representing causal knowledge, as mathematical accounts for learning. [[18]](#footnote-17) Constructivism places more emphasis on manipulations of higher level concepts and ideas that are not directly present in observed data. For example, one might abstract the category of “solid objects” which is formed from many individual objects (but distinct from all of them). As the Colunga and Smith paper calls attention to, language could be intimately related to how such non-observed quantities can be operationalized in cognition. The right evidence can update theories in broad and general ways, as opposed to the Hebbian localized connection updates found in associative learning. Two of the most prominent such models are Causal Bayes Net (also referenced as graphical models or belief networks) and Hierarchical Bayesian Models. There are numerous variations of these two models; the specific manifestations relevant to the conducted experiment are described in more detail below.

# Experiment - Forming Causal Inductive Biases

In this section we assess how well a constructivist model does in forming accurate inductive biases in a causally suggestive environment. While the theoretical and psychological underpinnings of rational constructivism are quite compelling, the theory lacks extensive computational exploration. Existing work on top-down probabilistic models have focused on their conceptual implications and empirical correspondence to human behavior. Our study focuses specifically on Hierarchical models and their sample efficiency, disregarding discussions of psychological or neurological plausibility. Furthermore, we explore how the generalizability of causal models varies under different data-generating procedures. Simulation makes it far easier to control the ground truth data generation process, which is either unknown or fixed in the real world. Approaching these models from a normative rather than descriptive lens, our goal is to assess how the model behaves when presented with different artificial datasets to probe the influence of the environment on the formation of inductive biases.

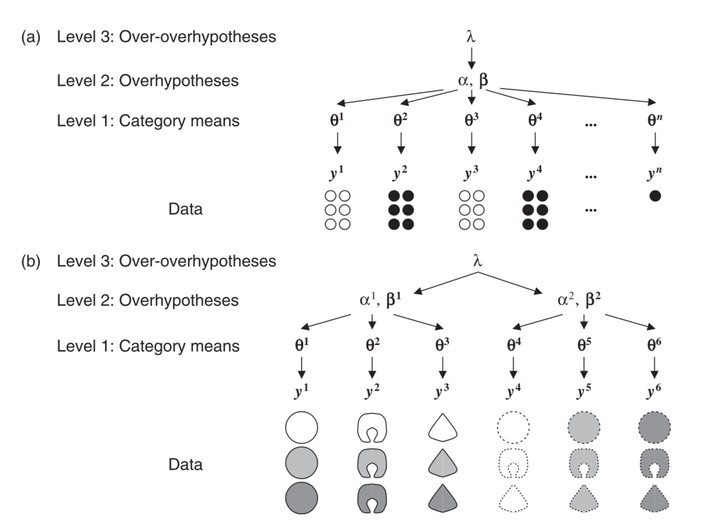
The toy task that has been set up centers around identifying whether a given fruit is poisonous. We assume that fruits only have three discernible attributes: color, texture, and shape. These attributes are assumed to be fixed constraints of the model’s perception system; our model’s job is not to identify which features exist but to learn which kinds of fruit are poisonous based on its features. At the same time, it must construct an overhypothesis about which features are relevant for whether or not a fruit is poisonous. For simplicity, we assume all of these features are discrete. Fruits observed in the environment can be one of four colors (red, purple, blue, green); 3 textures (smooth, spiky, rough); and 3 shapes (sphere, cube, star). When generating n data points, n fruits with randomized features are created. For now, there is a uniform distribution on which kinds of fruits exist. The label for each fruit is determined using a specified function is\_poisonous; this function determines the ground truth for the data-generation and is modified for different tests.

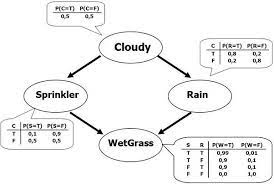


An example of a data set of 10 fruits generated by the described procedure is shown here. In this case, spherical fruits are poisonous. .

The model must construct two tiers of knowledge from as few fruits as possible. First, it must learn that (in this case) spherical fruits are poisonous and second infer the overhypothesis that shape is the relevant property in defining whether or not a fruit is poisonous. Technically, our formulation is similar to any supervised classification task with multiple nonbinary discrete features and a single binary label; in the actual code, the features and labels are encoded as integers anyway.

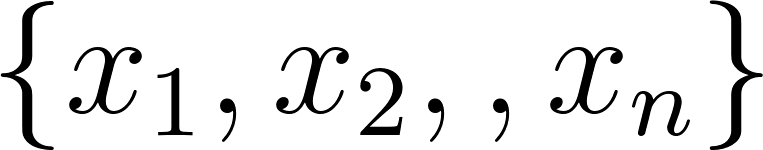
### Methodology (consistent aspects across experiments)

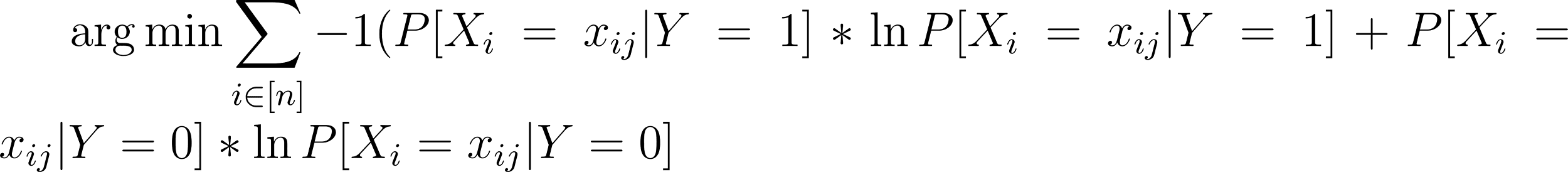
Kemp et al explore how Hierarchical Bayesian Models (HBM) can learn overhypotheses at multiple levels of abstraction. [[19]](#footnote-18) In his illustrative example, he describes the task of guessing the color of the next marble to be drawn from a bag (image from Kemp et al). On top of the data itself could be the subjective probability estimates for the observed bags. After viewing multiple bags that are all black or all white, one might form the overhypothesis that the bags are homogenous in color. And after seeing multiple bags, one might form an over-overhypothesis about the distribution of colors among the bags in general. The second level allows one to generalize the contents of the unseen marbles in a known bag. The third level would allow one to guess the contents of an unseen bag. HBMs are typically called “Bayesian” because the knowledge of each layer is represented by a probability distribution with some parameter whose value encodes the inductive bias or expectation associated with that layer. HBM’s learn from data by performing Bayesian updates on posterior probabilities. The prior of a given layer is directly related to the posterior of the immediately more abstract (next highest level) layer. 

The type of Hierarchical model implemented in our work draws inspiration from HBM’s but adopts different mechanisms that reflect the distinct task of forming causal models that encode how our world variables (shape, texture, color, poisonous) relate to each other. Such an overhypothesis cannot be neatly represented by a probability distribution. Instead, our Level 2 representation takes the form of a Bayes Net. Bayes Nets are Directed acyclic Graphs (DAGs), meaning they are represented by a finite set of nodes and directed edges. Each node represents a specific variable while each edge points from an alleged cause to the variable of the effect (image source: [[20]](#footnote-19)) . Edges also carry conditional probabilities conveying the extent of the causal connection, allowing for more nuanced representations. If two nodes are not connected, the DAG is implying they are independent. The task of learning the optimal causal graph for a given data set has been proven to be NP-Hard. [[21]](#footnote-20) This is intuitive if one considers that for n different variables or nodes, there are O(n!) possible graphs that can be constructed- this problem is worse than exponential! But is far less daunting to examine whether a given graph structure is sensible (making the problem polynomial-time verifiable). And the independence/ conditional independence tests that any filtering mechanism would rely on to rule out an edge between specific variables are only reliable if there are a *huge* number of samples. It is precisely due to these difficulties that learning causal Bayes Nets is an excellent case study for inductive bias. For without some means of constraining which graphs are considered, the problem is intractable. Because the problem is NP-Hard, the existing mechanisms for learning causal DAGs all make some sort of simplifying assumption or provide some justifiable heuristic for narrowing the search. Across the experiments we conduct, the Level One knowledge representation encoded in our hierarchical model stands for some directly observed or calculated aspect of the training fruit dataset while the Level Two representation is the DAG inferred on the basis of that direct observation. And if the mechanism of forming the Bayes Net relies upon implicit assumptions on what kinds of DAGs would be plausible, that would constitute an over-overhypothesis. Thus, the hierarchical model (HM) used in our experiments draws from both HBMs and Bayes Nets but is not equivalent to either of them. 

## Experiment One: Identifying the Causal Variable

Our first experiment establishes the ability of our basic HM to learn which variable is causally relevant in the simplest possible data-generating scenario. Our data ground truth is that there is only one causally relevant feature and prespecified value of that feature which makes a fruit poisonous. The dataset shown above was generated under precisely these assumptions; shape is the causally relevant property while “sphere” shaped fruits are poisonous. Fruit color and fruit texture have no bearing on whether or not a fruit is poisonous- but due to the fact that those attributes are randomly assigned, there is a high potential for spurious correlations in sparse datasets. For example, if every randomly generated spherical fruit *just so happens* to be purple, then a simple feature-label correlation measuring algorithm would fail to identify the causally relevant variable.

The algorithm chosen for selecting the DAG was a modified version of Cheng et al’s work, “Learning Belief Networks from Data: An Information Based Approach” [[22]](#footnote-21). Simulating conditions of data sparsity (one does not get to test very many poisonous fruits), seeing 5-10 fruits would be insufficient to use explicit conditional independence tests. Noise in the randomly sampled data would induce spurious correlations that would corrupt the results; with individual samples being hugely influential with a single-digit sample size and with 36 possible fruits, independent variables would seem dependent and vice versa. That is why an information theory based approach was taken. The original paper measures the mutual information between all variables; in this experiment we decided to apply a Level 3 over-overhypothesis that constrained which Bayes Nets were considered beyond that of the paper. Leveraging the fact that it was known that only the relationship between each of the fruit features and the poisonous label were decision-relevant, feature independence was assumed (essentially the Naive Bayes assumption) and a priori the number of possible Bayes nets was drastically reduced. Thus, the model selected the feature(s) with the lowest entropy on the conditional distributions on the label and constructed the DAG accordingly; the Level Two representation had access only to the Level One representations, not to the data itself. For features [](https://latex-staging.easygenerator.com/eqneditor/editor.php?latex=%5C%7Bx_1%2C%20x_2%2C%20%E2%80%A6%2C%20x_n%5C%7D#0) select the feature that minimizes the sum of the entropies conditioned on the poisonous and not poisonous distributions.

[](https://www.codecogs.com/eqnedit.php?latex=%5Carg%5Cmin%20%5Csum_%7Bi%20%5Cin%20%5Bn%5D%7D%20-1(P%5BX_i%20%3D%20x_%7Bij%7D%20%7C%20Y%3D1%5D%20*%20%5Cln%20P%5BX_i%20%3D%20x_%7Bij%7D%20%7C%20Y%3D1%5D%20%2B%20P%5BX_i%20%3D%20x_%7Bij%7D%20%7C%20Y%3D0%5D%20*%20%5Cln%20P%5BX_i%20%3D%20x_%7Bij%7D%20%7C%20Y%3D0%5D%20%20#0)

Where i refers to which feature (shape, texture, color) and j refers to which value (red, blue, green)

## Experiment Two: Altering Complexity of Ground Truth Labels

### Experiment 2a: (Implemented)

This experiment was partially implemented, but does not have results. In Experiment One, the data generating process was based on a singular deterministic feature value (spherical fruits → shape) which was randomly regenerated each time. This was in line with the Naive Bayes independence heuristic, which only hypothesized DAGs matching this deterministic feature value. Having successfully created a model that was able to identify both the specific value and abstract the causal feature responsible for poisonous fruits, we now investigate how the same model performs when there is not such a convenient correspondence between inductive bias and reality. The same information-entropy based method is used, except now the ground truth reality is more complicated. The learning algorithm picks the feature with the lowest entropy (and only identifies two causal mechanisms if the entropies of multiple variables happen to be the same). What if there are multiple types of poisonous fruits?

### Experiment 2b: (Not implemented:)

What if the Naive Bayes assumption was incorrect? Instead of randomly selecting fruits, what if the features were correlated- “Red fruits are more likely to be cube shaped than blue fruits”. What is the change in behavior and sample efficiency? After all, our assumption was centered around the fact that inter-feature dependencies were not as relevant to the task of predicting whether a fruit was poisonous or not. Since in this experiment, the label is still generated deterministically based on one feature, (still only spherical fruits being poisonous, as opposed to Experiment 2a) it is unclear how well the HM will do.

## Experiment Three: Probabilistic Bayes Nets with Associative Learning

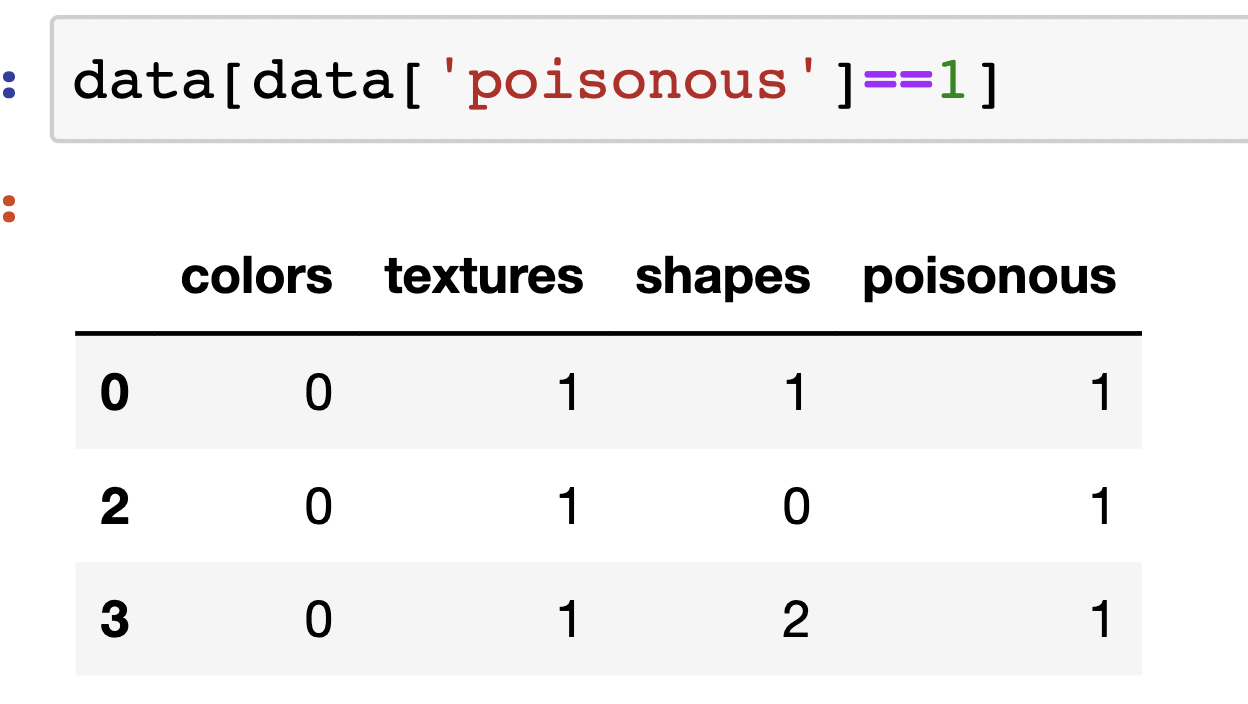
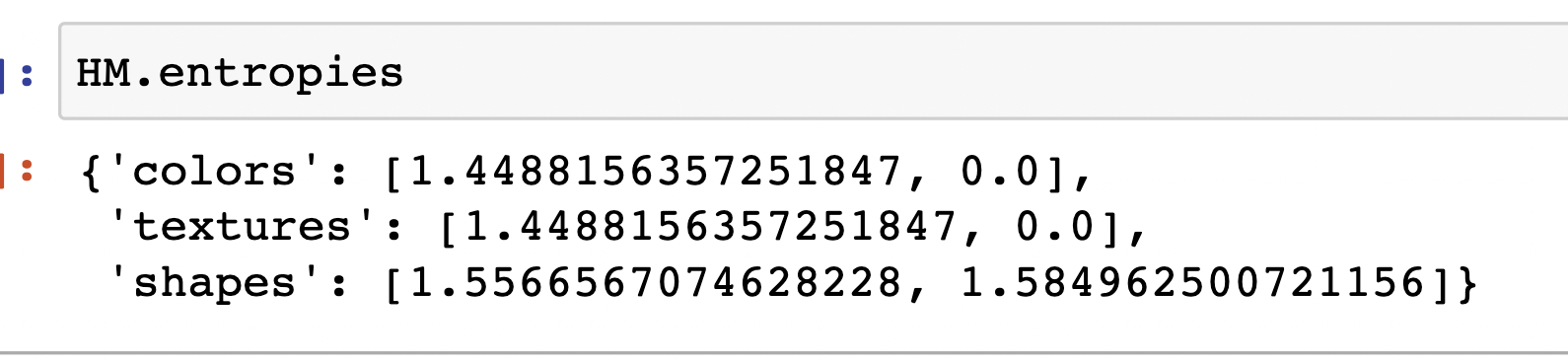
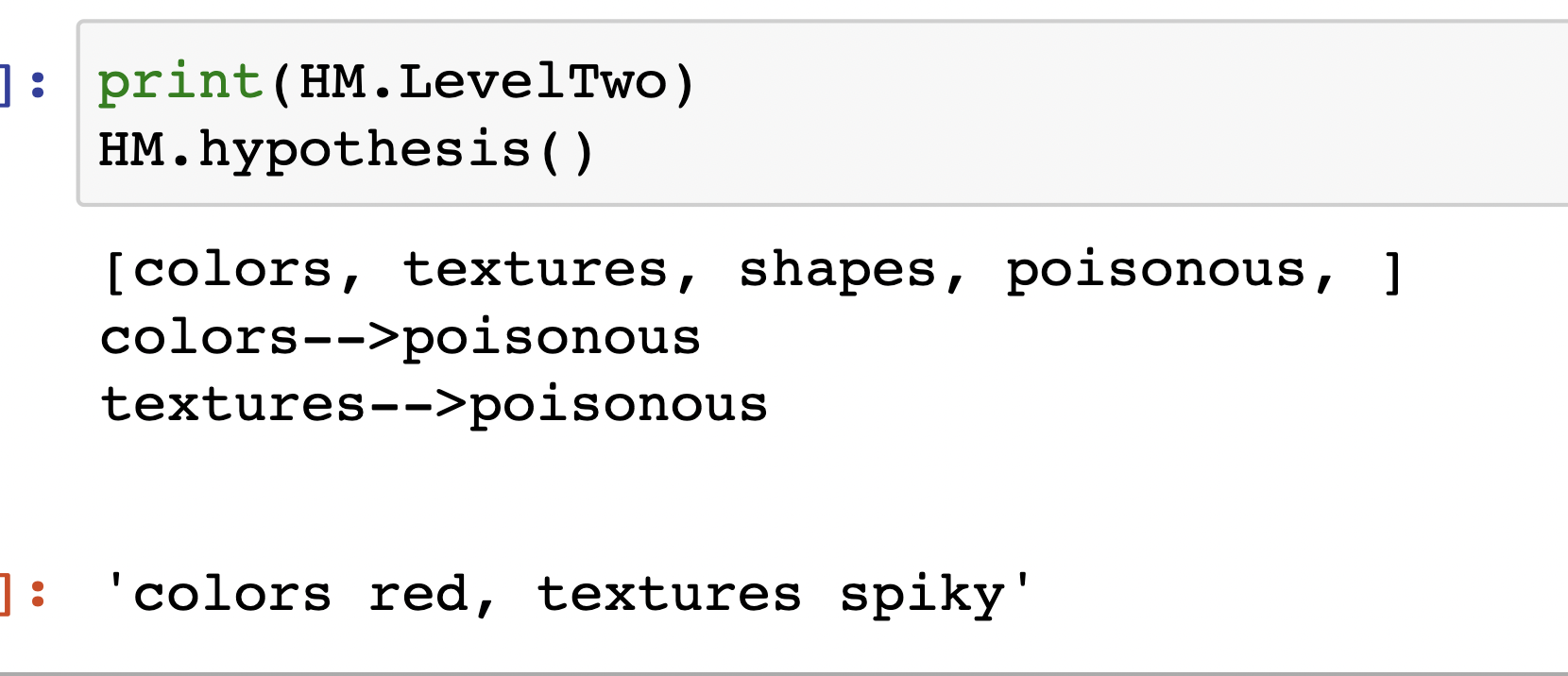
This experiment was not implemented, but is left as a future work. In the first two experiments, the causal model was not a true Bayes Net because the connections were not weighted by the strength of the causal connection. This made it somewhat unwieldy for testing generalizability. Implementing an associative mechanism where instead of simply selecting the feature with the minimum entropy, connections are strengthened in some approximation of Hebbian learning. These values could very easily be normalized using a softmax function to be a valid probability distribution. The selection of the causal feature in the is\_poisonous() function would be modified from being randomly reselected (sampled from a uniform distribution) every trial to sampled from a categorical Bernoulli distribution (weighted so that certain features are causal more often than others). This way, the Hebbian connections can aim towards learning a Level 3 Overhypothesis about which features are causally relevant across different data sets. This would mimic learning something similar to the shape bias through repeated associations with different kinds of objects. There are several key advantages to this setup. Firstly, it helps deal with the issue of ambiguous data by replacing hard assignments of causal structures with estimates that strengthen or weaken with time. It also allows for more graded causal hypotheses. It is likely that real world data is never uniqely determined by a single property but arises as the result of interactions between multiple cues of varying strength.

More complicated is\_poisonous functions involving two variables?

# Results and Discussion/Conclusion

## Experiment 1

There arose situations (as shown in the above pictures) where every randomly generated poisonous fruit had exactly the same set of features (this trivially occurs if there is only one poisonous fruit generated) In general, (somewhat surprisingly), our HM was fairly robust to ambiguity. The total feature entropy was based on not just the poisonous fruits (which were sometimes indistinguishable) but also the non-poisonous fruits. Even in the case of spurious ambiguity, there tended to be less entropy for the causally relevant variable because the cause of poison was *never* present in non-poisonous fruits, guaranteeing a zero-value conditional probability that lowered the overall entropy.

* + 
* 
* 

Above is a walkthrough of how the model assessed an interesting edge case in the data where it did not succeed. The ambiguity in which features mattered in this case arose because poisonous fruits were always red and spiky. Conditioning on non-poisonous fruits in this case did not work because the distribution P[Xi|Y=0] *by chance* was the same for both texture and color. Thus, by the information entropy metric, the two features were identical. Since the two features were genuinely indistinguishable, the HM formed a causal DAG with two edges; one for color and one for texture. This would be marked as incorrect by the accuracy metric (the underlying rule was that red fruits were poisonous) but the HM arguably made a smarter generalization with the provided data.

Since our limited setup was not very conducive to testing generalization (unseen data would either match or not match the deterministic edge matched; the new data would have a completely new value) due to the deterministic nature of the created DAGs, we stuck to evaluating sample efficiency. How quickly can we learn the underlying rule? We trained the model on a variable number of fruit samples; on the x-axis, we have the number of fruits the model learned from and on the y axis we have accuracy (percentage successful) over 100 trials. A “successful” trial was one in which both the feature value and the causal structure were correctly ascertained. Although generalization is usually tested by separating data into training and testing sets, in our simulation it is unnecessary because we have direct access to the data-generating ground truth. A model that has learned and acts upon the correct causal structure has automatically generalized well.

Although this figure graphs the total number of fruits seen, it ended up being the case that the number of *poisonous fruits generated* was more relevant. If not a single poisonous fruit was generated, the model would justifiably fail to learn anything meaningful. And if only one poisonous fruit was generated, it would be extremely hard to disambiguate which property was responsible- even factoring in the non-poisonous entropies, it would be somewhat of a shot in the dark. Of course, since a given randomly generated fruit has some fixed probability of being poisonous, the quantities are extremely correlated. It was primarily at around 6-7 fruits that 1-2 poisonous fruits would reliably pop up and the model was able to achieve accuracy above 80%. 20 fruits were sufficient for near-perfect accuracy (0.98) which leveled to perfect accuracy by the time 30 fruits were viewed.

| Number of Fruit Samples | Accuracy over 100 trials |
| --- | --- |
| 1 | 0.0 |
| 3 | 0.28 |
| 5 | 0.57 |
| 6 | 0.65 |
| 8 | 0.87 |
| 10 | 0.9 |
| 15 | 0.94 |
| 20 | 0.97 |
| 30 | 1.0 |

Interestingly, the HM was also tested specifically on data points fitting the pattern tested by Gopnik in the difference making study. If our HM sees three data points that explicitly point out a switch from not-poisonous to poisonous by the flipping of a singular variable, it is able to induce that that is the causal variable most of the time. Therefore, our HM also fits the data of Gopnik’s study.

### 

### Code available at <https://github.com/SammyAgrawal/COGS4800>

## 

## Bibliography

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*(6), 1085–1108. [https://doi.org/10.1037/a0028044](https://psycnet.apa.org/doi/10.1037/a0028044)

Colunga, E., & Smith, L. B. (2005). From the Lexicon to Expectations About Kinds: A Role for Associative Learning. *Psychological Review, 112*(2), 347–382. [https://doi.org/10.1037/0033-295X.112.2.347](https://psycnet.apa.org/doi/10.1037/0033-295X.112.2.347)

Goddu, M. K., & Gopnik, A. (2020). Learning what to change: Young children use “difference-making” to identify causally relevant variables.*Developmental Psychology, 56*(2), 275–284. <https://doi.org/10.1037/dev0000872>

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. Developmental Science, 10, 307–321. doi:10.1111/j.1467-7687.2007.00585.x

Griffiths, Thomas L., et al. "Probabilistic models of cognition: Exploring representations and inductive biases." *Trends in cognitive sciences* 14.8 (2010): 357-364.

Cheng, Jie, David A. Bell, and Weiru Liu. "Learning belief networks from data: An information theory based approach." *Proceedings of the sixth international conference on Information and knowledge management*. 1997.

Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review, 126*(6), 841–864. [https://doi.org/10.1037/rev0000153](https://psycnet.apa.org/doi/10.1037/rev0000153)

### Citations:

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131

Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review, 126*(6), 841–864. [https://doi.org/10.1037/rev0000153](https://psycnet.apa.org/doi/10.1037/rev0000153)

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review, 118*(1), 110–119. [https://doi.org/10.1037/a0021336](https://psycnet.apa.org/doi/10.1037/a0021336)

Woodward, J. The problem of variable choice. *Synthese 193, 1047–1072 (2016)*. <https://doi.org/10.1007/s11229-015-0810-5>

1. White, Jennifer C., and Ryan Cotterell. "Examining the inductive bias of neural language models with artificial languages." *arXiv preprint arXiv:2106.01044* (2021). [↑](#footnote-ref-0)
2. Morrison, Katelyn, et al. "Exploring corruption robustness: Inductive biases in vision transformers and mlp-mixers." *arXiv preprint arXiv:2106.13122* (2021). [↑](#footnote-ref-1)
3. Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. Developmental Science, 10, 307–321. doi:10.1111/j.1467-7687.2007.00585.x [↑](#footnote-ref-2)
4. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131 [↑](#footnote-ref-3)
5. Moruzzi, C. (2022). Toward Out-of-Distribution Generalization Through Inductive Biases. In: Müller, V.C. (eds) Philosophy and Theory of Artificial Intelligence 2021. PTAI 2021. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 63. Springer, Cham. https://doi.org/10.1007/978-3-031-09153-7\_5 [↑](#footnote-ref-4)
6. OpenAI. GPT-4 Technical Report. (2023) [↑](#footnote-ref-5)
7. Colunga, E., & Smith, L. B. (2005). From the Lexicon to Expectations About Kinds: A Role for Associative Learning. Psychological Review, 112(2), 347–382. [https://doi.org/10.1037/0033-295X.112.2.347](https://psycnet.apa.org/doi/10.1037/0033-295X.112.2.347) [↑](#footnote-ref-6)
8. Barbara Landau, Linda B Smith, Susan Jones, Syntactic context and the shape bias in children's and adults' lexical learning, Journal of Memory and Language, Volume 31, Issue 6, 1992, Pages 807-825, ISSN 0749-596X,

   https://doi.org/10.1016/0749-596X(92)90040-5. [↑](#footnote-ref-7)
9. Tek S, Jaffery G, Swensen L, Fein D, Naigles LR. The Shape Bias is Affected by Differing Similarity Among Objects. Cogn Dev. 2012 Jan;27(1):28-38. doi: 10.1016/j.cogdev.2011.09.009. PMID: 22247590; PMCID: PMC3255290. [↑](#footnote-ref-8)
10. Woodward, J. The problem of variable choice. Synthese 193, 1047–1072 (2016). <https://doi.org/10.1007/s11229-015-0810-5> [↑](#footnote-ref-9)
11. Goddu, Mariel K., and Alison Gopnik. "Learning what to change: Young children use “difference-making” to identify causally relevant variables." Developmental psychology 56.2 (2020): 275. [↑](#footnote-ref-10)
12. Xu, Fei, Kathryn Dewar, and Amy Perfors. "Induction, overhypotheses, and the shape bias: Some arguments and evidence for rational constructivism." The origins of object knowledge (2009): 263-284. [↑](#footnote-ref-11)
13. Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*(6), 1085–1108. [https://doi.org/10.1037/a0028044](https://psycnet.apa.org/doi/10.1037/a0028044) [↑](#footnote-ref-12)
14. Xu, F. (2019). Towards a rational constructivist theory of cognitive development. Psychological Review, 126(6), 841–864. [https://doi.org/10.1037/rev0000153](https://psycnet.apa.org/doi/10.1037/rev0000153) [↑](#footnote-ref-13)
15. Griffiths, Thomas L., et al. "Probabilistic models of cognition: Exploring representations and inductive biases." *Trends in cognitive sciences* 14.8 (2010): 357-364. [↑](#footnote-ref-14)
16. Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*(6), 1085–1108. [https://doi.org/10.1037/a0028044](https://psycnet.apa.org/doi/10.1037/a0028044) [↑](#footnote-ref-15)
17. Hopfield, John J. "Neural networks and physical systems with emergent collective computational abilities." *Proceedings of the national academy of sciences* 79.8 (1982): 2554-2558. [↑](#footnote-ref-16)
18. Griffiths, Thomas L., et al. "Probabilistic models of cognition: Exploring representations and inductive biases." *Trends in cognitive sciences* 14.8 (2010): 357-364. [↑](#footnote-ref-17)
19. Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. Developmental Science, 10, 307–321. doi:10.1111/j.1467-7687.2007.00585.x [↑](#footnote-ref-18)
20. Image: <https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e> [↑](#footnote-ref-19)
21. D.M. Chickering, D. Geiger, D. Heckerman, Learning Bayesian networks is NP-hard, Technical Report, MSR-TR-94-17, Microsoft Research, Microsoft Corporation, 1994. [↑](#footnote-ref-20)
22. Cheng, Jie, David A. Bell, and Weiru Liu. "Learning belief networks from data: An information theory based approach." *Proceedings of the sixth international conference on Information and knowledge management*. 1997. [↑](#footnote-ref-21)