Category Explanations

Fred Callaway

September 19, 2014

My goal for this summer was to conduct an exploratory investigation of explanation by appeal to category. My guiding hypothesis was that explanations using "because" take their meaning from causality, and that causality in turn takes its meaning from counterfactuals. For example, "Ferraris are fast because they are sports cars" means *Ferraris wouldn't be fast if they weren't sports cars.* This account of explanation is desirable because it can parsimoniously explain many different types of explanations that have previously been thought of as different: mechanistic, teleological, and of course, categorical. Although category explanations are relatively uncommon compared to these other kinds of explanation, they are an interesting subject of inquiry because they do not have an intuitive causal meaning—few people would say that being a sports car causes a Ferrari to be fast. [1]

There were two main sections to my work, the first focused on experimentation, and the second focused on Church modeling. Unfortunately, these two strategies did not interact significantly, due mostly to the inherent clash between the complexity of real-world category structure and the simplicity required by Church models. The exception is that, with both strategies, I focused on the super-category as an important constraint in considering counterfactual alternatives. In the first of the next two sections, I will describe how I discovered the necessity for a constraint on alternatives, and how I arrived at the super-category as a suitable source of this constraint. In the second, I will describe how this constraint emerges naturally from a counterfactual model in Church. I will also discuss the specific strategies I used and difficulties I encountered while following each of these two paths.

Experimental work

In the first five weeks, I ran a total of six mTurk experiments with steadily increasing success. These experiments were guided by a theory that category counterfactuals are interpreted in terms of a super-category. In the next section

[1] This brings up the interesting question of whether causation and explanation could be distinct phenomena that both take their meaning from the counterfactual (as opposed to the theory I initially presented that explanation takes its meaning from causation, which in turn takes its meaning from counterfactuals). I will remain agnostic to this question in this report, and simply discuss the link between counterfactuals and explanations.

I describe the process leading to this theory and its formalization into a model. Then I describe the experiments I undertook to test the model. Overall, this phase of my work did not generate very useful results. However, I am inclined to attribute this to my own inexperience in experimental design and modeling as well as the inherent noisiness of subject performance, rather than to a shortcoming of the theory. That being said, the theory is certainly not complete, something I discuss further in the Next Steps section.

Creating a model

All my experiments were guided by a model that I took from Gerstenberg et al. (2012) that predicted explanations in a physical domain. I simply substituted physical events with features and categories to produce the following model:

p("feature because category")$\propto$p(feature|category) - p(feature|$\neg$category).[2]

A problem presents itself, however: the problem of alternatives. In the physical situations modeled by Gerstenberg et al., there are two clearly defined possibilities: one in which ball A is present, and one in which ball A is absent. With categories, however, there is a huge number of possible alternatives. This problem comes to force when we consider two competing explanations: "The animal has feathers because it is a peregrine falcon" and "The animal has feathers because it is a bird." In both cases, p(feature|category) is very high and p(feature|$\neg$ category) is very low. Although there are slightly more feathered things that are not peregrine falcons than that are not birds, the immense number of non-bird categories overshadows this small, but crucial, difference. The model fails to predict that the second of the two explanations is considerably superior.

We must constrain the alternatives. Because the problem in our example arises from the huge number of non-bird categories, I decided to factor these out by only considering categories within the same super-category as viable alternatives. In our example, we would consider falcons that are not peregrine falcons and animals that are not birds. With this constraint, the percentage of feathered beings in each group is now very high and very low, respectively. When we subtract these values from p(feathers) for peregrine falcons and birds respectively (both near 1.0), we get a higher result for the "bird" explanation. Thus, the second explanation takes its rightful place as the better of the two. Incorporating this constraint into the original equation, we have:

p("feature because category") $\propto$ p(feature|category) - p(feature|super-category, $\neg$category).

[2]This equation predicts negative probabilities, but in line with Gerstenberg et al., we can interpret a negative value as predicting "category prevented feature" or "$\neg$feature because category." Alternatively we could normalize to generate values between 0 and 1.

Experimental design and results

I ran a total of six mTurk experiments testing the model described above, but I will only discuss four of them here. Example stimuli for all experiments are included in the appendix.[3] In all these experiments, the general strategy was to ask subjects which category best explained a feature, and compare the results to model predictions based on priors—p(feature|category) and p(feature|¬category)—extracted from a different set of subjects. In **Experiment 1A,** subjects were presented with a statement and four possible explanations appealing to multiple systems and levels of categories. Subjects ranked each explanation with sliders. In **Experiment 1B**, subjects were asked what percentage of members in a category had some feature. I approximated p(feature|super-category, ¬category) as simply p(feature|super-category). Ultimately, I was only able to analyze the two explanation types that I could get both priors for.

Model predictions correlated only slightly with experimental data when analyzing both explanation types together. Separating by explanation type, I found that the super-category term only improved predictions for sub-type explanations (e.g. sports cars, evergreens). The p(feature|category) term made better predictions by itself for mid-type explanations (e.g. boats, cars, mammals). These results provided hope for my theory, but they were too weak to make a convincing argument with. I attributed this weakness to complexity of the experiment and the poor approximation of the counterfactual prior.

I addressed these issues in a second round of experiments. In **Experiment 2A**, subjects saw an image with a prompt asking why the pictured object had some feature and a leading *"Because it is a. . . "*. Subjects finished the sentence with one of three possible category levels from a single taxonomic branch. I used forced choice rather than rankings because it is closer to what we must do in every day speech and thus, may provide a better window into subconscious knowledge. In **Experiment 2B**, I elicited priors with a guessing game design, thinking that it might access subconscious knowledge better than an explicit statistical question. Subjects saw a prompt of the form *"I'm thinking of a* CATEGORY*. What's the chance that it* HAS FEATURE*?"*

Overall, predictions correlated with subject responses better than in Experiment 1. The simple p(feature|category) prior correlated much more strongly, confirming my hypothesis that simplifying the experiment would improve results. For this analysis, I used linear scaling to normalize the model predictions so that predictions for each possible response to a single question summed to 1. The model performance is shown in figure 2 with a correlation of $r^2 = .23$. Again, however, the counterfactual term only improved predictions for sub-type explanations, actually lowering correlation for the other two response types. I

[3] The experiments themselves can be viewed online at:
Exp 1A: http://cocolab.stanford.edu/experiments/explanation/experiments/cat2/cat2.html
Exp 1B: http://cocolab.stanford.edu/experiments/explanation/experiments/cat3/cat3.html

Exp 2A: http://cocolab.stanford.edu/experiments/explanation/experiments/cat5/cat5.html
Exp 2B: http://cocolab.stanford.edu/experiments/explanation/experiments/cat6/cat6.html

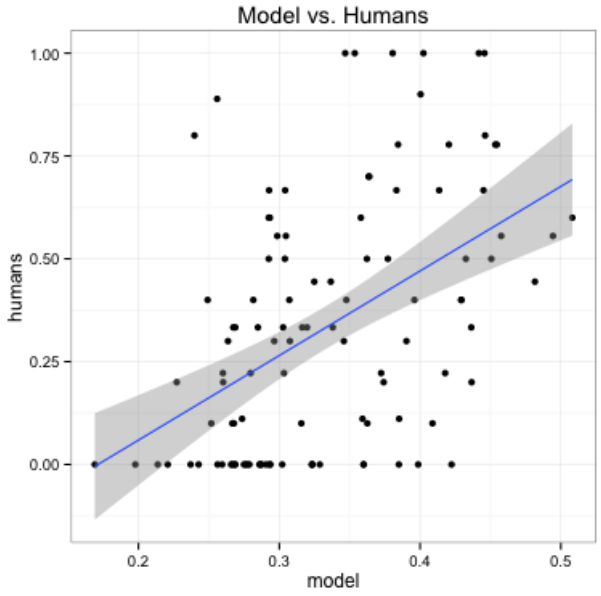


Figure 1: Model performance in Experiment 2

do not have a good explanation of this phenomenon.

In retrospect, I think the guessing game design may have actually introduced more confounds than it eliminated. In particular, there were likely pragmatic effects associated with the communicative goal inherent to the task. Furthermore, if priors are indeed as variable as they appear in the data, it is problematic to use the priors of one subject to explain the behavior of another. Finding an effective way to elicit priors is critical if we wish to study naturalistic category explanations, a task essential to forming a complete theory; however, we can avoid this difficulty and still form a strong theoretical base using artificial category paradigms with experimentally controlled priors. This is the goal of the experiment I am currently working on.

Church modeling

I spent the second three weeks working with the new counterfactual church model in an attempt to create a working model of category explanations. Most of this time was spent troubleshooting various issues that emerged using the counterfactual model with complex generative models. Here I present a brief summary of the problems I encountered and the strategies I used to overcome them. I present a more detailed discussion of the model predictions and provide

a theoretical analysis of the model. [4]

Algorithmic obstacles

The first issue to arise was poor predictions for generative models with indirect causation. This was a result of causally linked variables being resampled independently.This results in bad predictions; for example, if a causes b and b causes c, the literal listener does not interpret "c because a" as implying that b is true. It is easy to see how this would be a problem when modeling hierarchical category structure. Andreas and I were able to solve this problem using exogenous randomness (Pearl, 2000) The result is that a variable a is likely to change its value if a variable b that a is dependent on changes. a will change its value with probability $\left|p(a|b) - p(a|\neg b)\right|$. For deterministic dependency, a will change whenever b changes; for very weak dependencies, changing b is unlikely to result in changing a.

The second problem I encountered was that of efficiency. Enumeration query is intractable with complex models because the complexity is exponential with respect to the number of variables—even a very simple model with eight variables takes twelve minutes to run. Rejection query avoids this problem but faces the new problem of extremely unlikely condition statements that arise in the counterfactual queries. Adding additional queries for pragmatic speakers and listeners quickly worsens this problem because a difficult-to-satisfy condition may have a query with many of its own difficult-to-satisfy conditions. This makes a full pragmatic model *very* slow. No clear solution has presented itself to me. Enumeration query is inherently intractable for large models; thus it not likely to provide the solution. MCMC query does not seem to run faster than rejection query and it provides even noisier results. One possibility is a modified rejection query that uses heuristics to avoid very low probability conditions. Unfortunately, I do not have the background in computer science to provide a fully analysis of this problem, let alone a solution.

Model predictions

Despite these difficulties, I was able to create a model that makes intuitively appealing predictions. This model can be found at forestdb.org/models/category-explanations. In the simple two-level category structure depicted in figure 2, the model predicts a super-category explanation (north) when both of its sub-categories are likely to have a feature (stripes), but a sub-category explanation (wug) when only one sub-category is likely to have a feature. The model predicts a super-category explanation roughly in proportion to the probability of the second sub-category (del) having the feature.

[4] A more complete description of the implementational difficulties can be found at: forestdb.org/models/exogenous-counterfactuals.

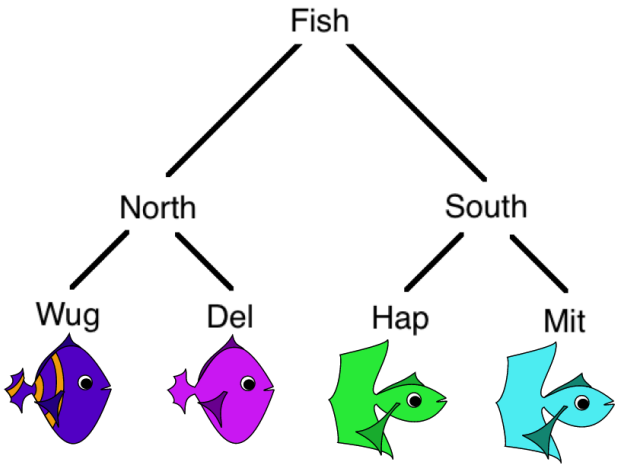


Figure 2: Example category structure

Model analysis

Although I do not yet have human data to compare model predictions to, I am not very optimistic about its performance. Its predictions, while somewhat appealing, do not have a strong basis. For example, the super-category effect in the current model arises because it is more likely to counterfactually resample within the super-category than into a different super-category—this is good. However, the effect is due to incidental properties of the category structure. It is possible to arrive at a sister sub-category by resampling at the sub-category or super-category level, while one can only arrive at a non-sister sub-category by resampling at the super-category level; thus, a greater proportion of possible resamplings results in the sister-subcategory. Incidental effects are often seen as indications of a models strength; however, in this case, the effect is dependent on intuitively unimportant characteristics of the category structure. Namely, changing the category priors to make the alternative super-category more probable weakens the super-category effect because resampling the super-category is less likely to result in its original value. It is unclear whether category priors should have any effect at all in this way, but it certainly should not so easily outweigh the super-category effect. One possible solution to this problem is to use a non-uniform sampling rate. Specifically, higher level categories should resample with less frequency. In fact, we could make this a general rule: variables further up in the causal chain should resample with less frequency. This is an intuitively appealing characterization of counterfactuals, and it also explains why people generally tend to choose explanations as causally "close" as possible.

Another way we could achieve this effect would be to counterfactualize not on categories, but on features. Similarity is generally higher between more closely related categories, thus changing a super-category would result in more feature

changes than changing a sub-category. In turn, if the shadow-object tends to maintain the original features, it will tend to remain in the same super-category independent of the category priors. This strategy has the additional benefit of assigning different resampling rates to categories at the same level. It is also appealing because it generates the super-category effect naturally as opposed to with an arbitrary rule. This feature-based strategy could be implemented by allowing the shadow model to randomly generate shadow-objects of all alternative sub-categories with equal probability. It would then noisily condition on the features of this new shadow-object being the same as those of the original object before returning the object to the shadow condition: being in a different category. This would result in shadow-objects that tend to be similar to the original object and thus tend to be in similar categories.

The feature-based strategy is more than an algorithmic trick. It represents a theoretical move towards representing categories as emergent phenomena of similarity structures rather than strict hierarchical trees. One could move further in this direction by eliminating the category-based generative model altogether and generating the possible shadow-objects using co-occurrence statistics of all possible features.[5] A category would be assigned to the object after the fact, and the shadow model would condition on this category being different from the original. Attempting to implement a model in this style and comparing its predictions to the category-based model is one of my long-term goals. In line with my general ideas about the interaction between explicit and emergent structures in cognitive systems, I think an ideal model will incorporate elements of both styles. This model would predict an effect of the high-level, explicit, category structure as well as the low-level, emergent, similarity structure.

On a similar line, the model currently forms feature-category connections entirely statistically, i.e. there is no explicit rule like "wugs have stripes." I suspect that people will not rely solely on statistical analysis, but will instead form simplifying rules. This predicts a sigmoidal transition from a "wug" to "north" explanation as p(stripes|del) increases and the rule shifts from "wugs have stripes" to "northern fish have stripes." Interestingly, the preliminary model results are suggestive of such a sigmoidal transition. I do not yet understand why this pattern emerges, nor am I certain the pattern will hold as I gather more data points. Explaining this pattern and incorporating explicit simplifying rules into the model are both part of my current work.

Next steps

My immediate goal is to run the artificial category experiment and compare the results to the model predictions. I will then use these results and the issues discussed in the previous section to guide my refinement of the model. I may also attempt to create more complicated category structures or look at effects

[5] Simply generating shadow objects with random features is a bad solution because it will generally result in objects that do not fit into any category. Categories can be viewed as emergent phenomena of cooccurrence statistics; thus, only creating objects with features that are likely to cooccur achieves a similar effect to creating objects based on explicit categories.

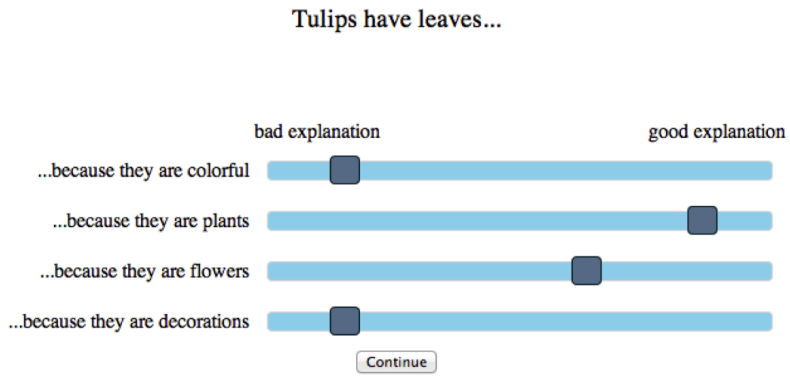
of other variables, such as category priors. I will likely go through many cycles of experimentation and model refinement to slowly increase the number of phenomena that I can account for. Throughout this process, I will be especially interested in the competing roles of explicit and emergent structure. In the more distant future, I may look at how people select between different category structures (Ross et al., 1999) when providing explanations. I may also look at the other kind of category explanation: explaining a category by appeal to feature. This is an interesting case because we get situations where explanation can go either way "it's a giraffe because it has a long neck" and "it has a long neck because it's a giraffe" seem to be equally acceptable. This is difficult from a causal perspective because causation is not a reciprocal relationship. One possibility is an elided "I know" before the 'other' kind, although this is not a very satisfying answer. Although I can no longer work full time on this project, there is a lot of interesting work to do, and I will devote as much time as I can to moving forward on it.

References

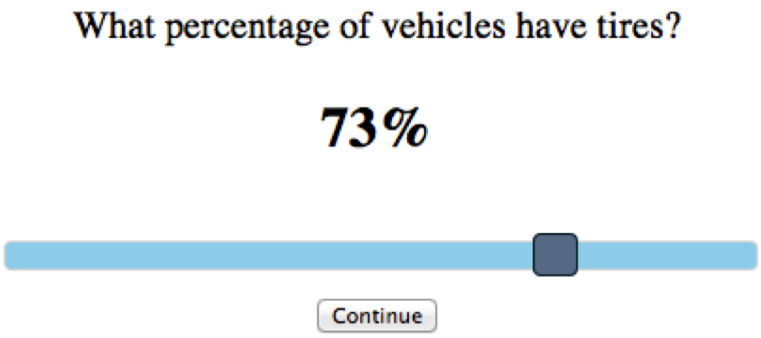
[1] Tobias Gerstenberg, Noah Goodman, David A Lagnado, and Joshua B Tenenbaum. Noisy newtons: Unifying process and dependency accounts of causal attribution. In *In Proceedings of the 34th*. Citeseer, 2012.

[2] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[3] Brian H Ross and Gregory L Murphy. Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive psychology*, 38(4):495–553, 1999.



(a) Experiment 1A



(b) Experiment 1B

Why does this thing have windows?

Because it's a(n) ...

- ○ vehicle
- ◉ car
- ○ sports car

Continue

(a) Experiment 2A

*"I'm thinking of **a plant** that's **not a tree**."*

What's the chance that it is alive?

very unlikely very likely

Continue

(b) Experiment 2B