Decision factors that support preference learning

Alan Jern and Charles Kemp
{ajern,ckemp}@cmu.edu
Department of Psychology
Carnegie Mellon University

Abstract

People routinely draw inferences about others' preferences by observing their decisions. We study these inferences by characterizing a space of simple observed decisions. Previous work on attribution theory has identified several factors that predict whether a given decision provides strong evidence for an underlying preference. We identify one additional factor and show that a simple probabilistic model captures all of these factors. The model goes beyond verbal formulations of attribution theory by generating quantitative predictions about the full set of decisions that we consider. We test some of these predictions in two experiments: one with decisions involving positive effects and one with decisions involving negative effects. The second experiment confirms that inferences vary in systematic ways when positive effects are replaced by negative effects.

Keywords: preference learning; decisions; probabilistic model; attribution

Suppose your friend Alice orders a boxed lunch that includes an eggplant sandwich and you are curious how much Alice likes eggplant sandwiches. The conclusion you reach could depend on several factors. If there were many other boxed lunches available, perhaps Alice's preference for eggplant sandwiches is relatively strong. If all boxed lunches except the eggplant sandwich box come with a free cookie, perhaps Alice's preference for eggplant sandwiches is extremely strong. On the other hand, if the eggplant sandwich is part of the only box that contains a cookie, perhaps Alice's preference for eggplant sandwiches is relatively weak and she really wanted the cookie. As these examples suggest, any given choice could potentially have many different explanations, and deciding which of these explanations is best is often a challenging inductive problem.

In cases like these, observing someone make a decision provides information about his or her desires or preferences. Two classic proposals along these lines are Jones's and Davis's (1965) correspondent inference theory of attribution and Kelley's (1973) ANOVA model, both inspired by Heider (1958). Both proposals identify some normative principles that predict when an observed decision provides strong evidence for an underlying preference. The ANOVA model has influenced subsequent computational accounts of learning and reasoning (Cheng & Novick, 1992), but there have been few computational accounts that address the issues emphasized by correspondent inference theory (see Medcof, 1990). Here we show that a simple probabilistic model captures some of the key principles of the theory, along with some additional principles not identified by Jones and Davis.

To explore the factors that support preference learning, we work with a space of what we call decision events—observed decisions among discrete choices. Each event involves a set of *options*, and each option may have one or more *effects*. For example, a restaurant may offer three boxed lunches (three options), and one of these lunches may include an eggplant sandwich and a cookie (two effects). One principle of correspondent inference theory asserts that unique effects are maximally informative: for example, if Alice chooses the only boxed lunch that includes an eggplant sandwich, perhaps her preference for eggplant sandwiches is relatively strong. A second principle asserts that as the number of chosen effects increases, the less strongly one can conclude that an actor sought one particular effect. For example, if Alice's choice happens to be the only box that contains an eggplant sandwich and the only box that contains a cookie, perhaps she likes the cookie rather than the eggplant sandwich.

Both of these principles, along with several others that we discuss, are captured by a simple probabilistic model known as the multinomial logit model (McFadden, 1973). This model is common in the economics literature, and has received some attention in the psychological literature (Bergen, Evans, & Tenenbaum, 2010; Lucas, Griffiths, Xu, & Fawcett, 2009). The model assumes that an actor assigns some utility to each effect, and chooses probabilistically among the options in proportion to the total utility assigned to each one. Given these assumptions, it is possible to work backward from an observed decision to infer the likely utility assigned to each effect. Lucas et al. (2009) showed that the model helps to explain how children use statistical information to make inferences about others' preferences (Kushnir, Xu, & Wellman, 2010). We build on this work and suggest that the model provides a comprehensive account of preference learning over the full space of decision events.

A space of decision events

The first step is to formally characterize the space of decision events. We will use a running example where an actor is given a choice between bags (i.e., options) that contain candies of different brands (i.e., effects). The actor chooses a bag containing a Brand x candy, and our goal is to infer the strength of the actor's preference for Brand x. Figure 1a shows the 14 distinct decision events that involve up to four effects. The event on the far left is a case where the choice set includes a single bag that contains only a Brand x candy; the event on the far right is a case where the choice set includes four bags each containing a candy from a different brand. Since the labels of the candies are not important in this example, a single representative is included for all decision events that are the same up to relabeling. For example, the event that involves a single bag containing x and a is equivalent to the event that

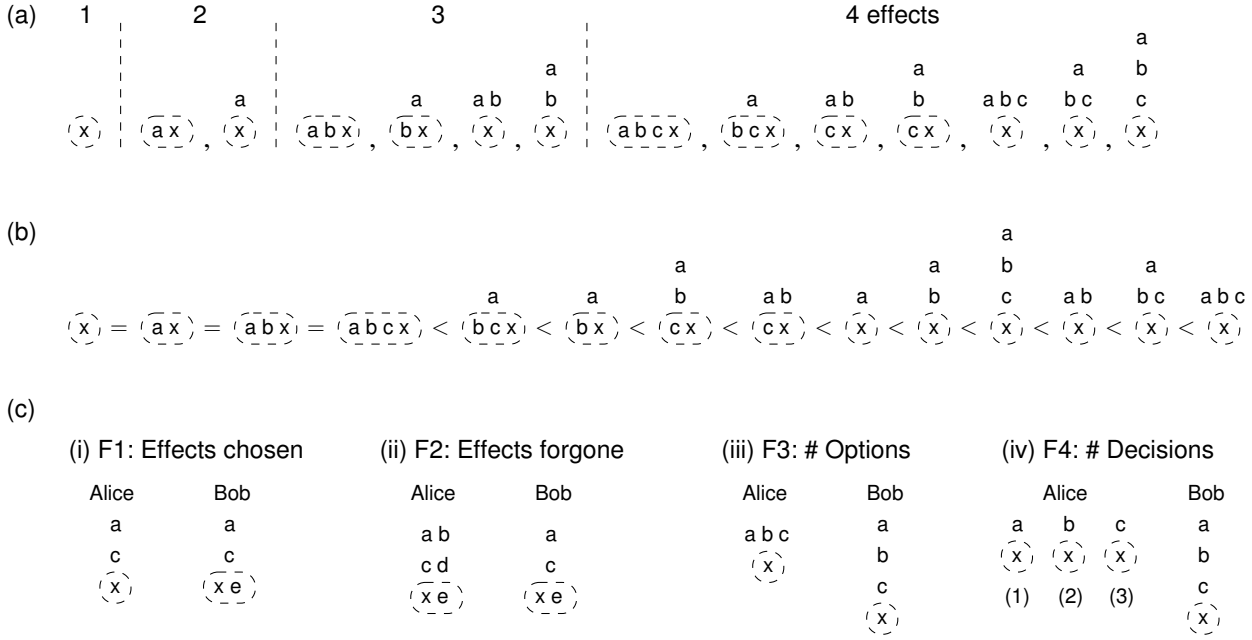

Figure 1: The space of decision events. Each decision is indicated by a set of options on each row. The effects of each option are listed as letters and the chosen option is circled. (a) The space of decision events with up to four effects, such that the option with effect x is always chosen. (b) These decisions can be ordered by how strongly they suggest a preference for x. (c) Four factors (F1–F4) that inform preference inferences. In each case, the two columns labeled Alice and Bob indicate the set of options each person was presented with. The four pairs of decisions illustrate cases in which Alice's and Bob's decisions differ in (i) the number of chosen effects, (ii) the number of forgone effects, (iii) the number of available options, and (iv) the number of decisions made.

involves a single bag containing x and b. The number of distinct events increases as the number of effects increases: for example, there are 26 distinct events that involve up to five effects, and 45 that involve up to six effects.

Given the full space of decision events in Figure 1a, it is natural to ask which events provide the strongest evidence that the actor likes Brand x. Figure 1b shows the ordering predicted by the model described in the next section when all effects are positive. None of the decision events at the far left is informative about a preference for x, since the actor is forced to choose a bag containing x if only one bag is available. The event on the far right provides strong evidence that the actor likes x, since she passed up a bag with three candies in order to acquire a single candy of Brand x.

A natural goal for behavioral research in this area is to establish an empirical ordering to compare with the predicted ordering in Figure 1b. We make a start in this direction by identifying several key factors that can be used to distinguish among events and studying the roles of these factors. Each decision event can be described in terms of three factors: the set of chosen effects, the set of forgone effects, and the distribution of forgone effects over the forgone options. These factors motivate the comparisons shown in Figures 1c(i)–(iii). Although Figure 1b focuses on events with up to four effects, note that the events in 1c include up to six effects to be consistent with our experiments, described later.

The first comparison (i) involves two events that differ only in the number of chosen effects. If asked to identify the actor with the greater preference for candy x, Alice seems to be the better choice, since Bob might have been interested in candy e. The second comparison (ii) involves events that differ only in the number of forgone effects. Here, Alice appears to have the stronger preference for x, since she passed up more effects in order to acquire an option that included x. The third comparison (iii) suggests that the distribution of forgone effects over the forgone options is also important. Both Alice and Bob passed up three effects, but Alice must have a strong preference for x if she chose a bag with one candy when she could have had a bag with three. So far we have focused on cases where a single decision is observed, but often we are able to observe an actor's behavior over time. The fourth comparison (iv) suggests that the number of decisions observed is relevant. Given that Alice chose Brand x on three separate occasions, perhaps she has a relatively strong preference for x.

In addition to considering each factor in isolation, comparisons between decision events may involve multiple factors. In some of these cases, two or more factors will support the same conclusion, but in others, some factors will be in conflict. Considering interactions of this kind is critical for developing a comprehensive account of preference learning. Thus, our experiments include two comparisons where factors F1

and F2 both apply.

Some of these factors have been previously studied. Newtson (1974) explored the influence of factors F1 (number of chosen effects) and F2 (number of forgone effects), and Lucas et al. (2009) also explored the role of factor F2. Factor F4 (number of decisions) is captured by Kelley's ANOVA theory, but to the best of our knowledge, factor F3 (number of options) has not been previously identified. In addition to exploring factor F3, we build on previous studies by demonstrating that the model described in the next section can account for the effects of all four factors.

A computational model of preference learning

In this section, we describe a simple formal model that helps to explain how observers make inferences about an actor's preferences. We assume the actor is presented with a set of n options $\{o_1, \ldots, o_n\}$, each of which produces one or more effects from the set $\{f_1, f_2, ..., f_m\}$. For simplicity, we assume that each effect is binary. Let u_i indicate the utility that the actor assigns to effect f_i, and suppose that the utility of each option is based on the utilities of the effects that it produces. The greater the utility of the option, the more likely the actor is to choose that option.

We make the standard assumption that utilities are additive. That is, if $\mathbf{f_j}$ is a binary vector indicating which effects are produced by option o_j and $\mathbf{u}$ is a vector of utilities assigned to each of the m effects, then the total utility associated with option o_j can be expressed as $U_j = \mathbf{f_j}^T \mathbf{u}$. We complete the specification of the model by applying the Luce choice rule (Luce, 1959), a common psychological model of choice behavior, as the function that chooses among the options on the basis of their utilities:

$$p(c = o_j|\mathbf{u},\mathbf{f}) = \frac{\exp(U_j)}{\sum_{k=1}^{n}\exp(U_k)} = \frac{\exp(\mathbf{f_j}^T\mathbf{u})}{\sum_{k=1}^{n}\exp(\mathbf{f_k}^T\mathbf{u})} \quad (1)$$

where c denotes the choice made.

Given these assumptions, we can use Bayes' rule to infer an actor's utilities after observing a choice he or she makes.

$$p(\mathbf{u}|c = o_j,\mathbf{f}) \propto p(c = o_j|\mathbf{u},\mathbf{f})p(\mathbf{u}) \quad (2)$$

In order to apply Equation 2 we must specify a prior on the utilities $p(\mathbf{u})$. We adopt a common approach that places independent Gaussian priors on the utilities: $u_i \sim \mathcal{N}(\mu,\sigma^2)$. For decisions where effects are positive, we set $\mu = 2\sigma$, which corresponds to a prior distribution that places approximately 2% of the probability mass below zero. Similarly, for negative effects, we set $\mu = -2\sigma$.

Experiment 1

We applied this model to a set of decision events designed to test the effects of the four factors in Figure 1c and compared its predictions to human judgments. Newtson (1974) previously studied factors F1 and F2; our first experiment replicated all of his conditions plus two more that focused on factors F3 and F4. Following Newtson, we examined the interactive effects of factors F1 and F2, including a case in which these two factors were in conflict.

Method

Participants 160 participants were recruited from the Amazon Mechanical Turk website. They were paid for their participation.

Design The experiment consisted of eight between-subject conditions, with 20 participants allocated to each condition. In each condition, participants read a story that described a pair of decisions that two people made. The full set of pairs is illustrated in Figure 2a. Each column of the figure represents a comparison between two decisions, shown at the top and bottom. The factors that each comparison manipulates are labeled above each column, where F1–F4 correspond to the factors in Figure 1c. In addition to manipulating each factor in isolation, we also considered two comparisons involving interactions between factors F1 and F2, shown in the last two columns.

Procedure Participants completed the experiment online. They were told that two people, Bob and Bill, were each given a choice between several bags of candy. The options were the bags of candy and the effects were different brands. As shown in Figure 2a(i), both Bob and Bill always chose the bag containing Brand x candy. Participants were then asked, "Based only on the above information, which person do you think likes Brand x candy more?" They provided their responses on a numerical scale from 1 (Bill likes Brand x candy more) to 8 (Bob likes Brand x candy more). The polarity of the scale was reversed for half of the participants. When varying the number of decisions (factor F4), participants were told that one person chose the bag with Brand x candy in it on three separate occasions.

Results

Model predictions For each decision event, we estimated the posterior distribution $p(\mathbf{u}|c = o_j,\mathbf{f})$ in Equation 2 using a Metropolis-Hastings sampler. We then computed the expected value of the utility for effect x, $E(u_x)$, by summing over all utilities except u_x. In order to produce predictions for the comparison involving factor F4, the number of decisions, we treated decisions as independent, such that $p(\mathbf{c}|\mathbf{u}) = \prod_i p(c_i|\mathbf{u})$. Because all the effects were intended to be clearly positive, we used a prior distribution on utilities with mean $\mu = 2\sigma$. The model predictions shown in Figure 2a(i) were based on a prior distribution with standard deviation $\sigma = 2$, but similar qualitative results were obtained with a range of variances.

The first row of Figure 2a(i) shows differences between the mean utilities for the actors in each pair. For each pair, the model predicts that the actor represented at the top of the plot has a greater utility for x. For every case, the model's predictions about the effects of the four factors are consistent with the intuitive explanations offered earlier for the comparisons in Figure 1c. This is also true for the case in which both

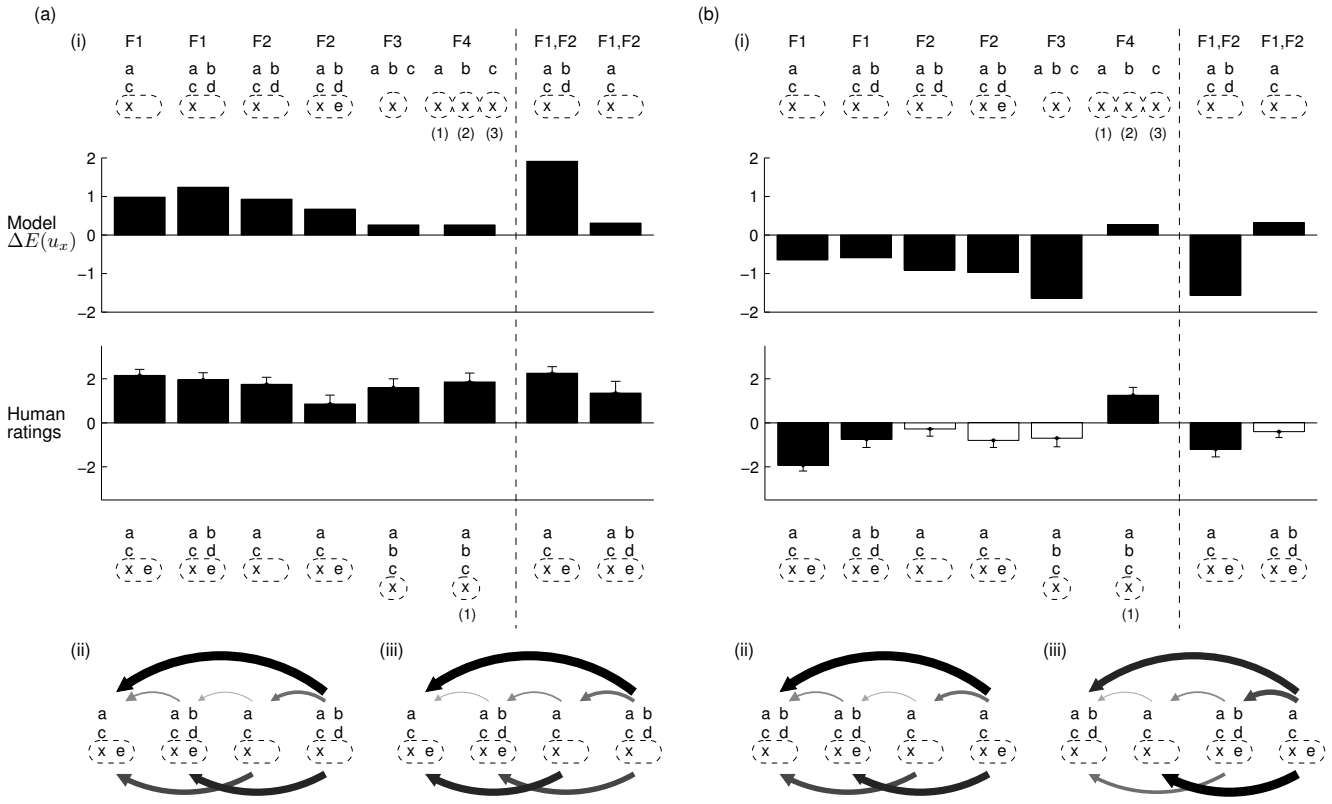

Figure 2: Model predictions and human data from (a) Experiment 1 and (b) Experiment 2. In each subfigure, panel (i) compares the effect magnitudes between the model and human participants for the eight studied cases. Error bars indicate standard errors. Black bars indicate results that are significantly different from 0. Panels (ii) and (iii) illustrate, for the model and human data respectively, an ordering by strength of attribution for the cases involving factors F1 and F2. The diagram also indicates by arrows the relative strengths of comparisons between the cases. The arrow is directed from the larger attribution to the smaller attribution and the thicker and darker the arrow, the larger the difference.

factors F1 and F2 are manipulated to support the same inference (the second column from the right), which leads to the model's largest predicted difference.

Of particular interest is the case where factors F1 and F2 are in conflict (the rightmost column). At first, it may not be clear which factor should carry more weight. The model, however, predicts that F1 should have a greater influence in this case. This prediction is a consequence of basic Bayesian inference, which implies that

$$\frac{P(\text{Bill loves x}|\text{Bill chooses bag 3})}{P(\text{Bob loves x}|\text{Bob chooses bag 3})} = \frac{P(\text{Bill chooses bag 3}|\text{Bill loves x})}{P(\text{Bob chooses bag 3}|\text{Bob loves x})}$$

where Bill is the actor who makes the choice at the top of Figure 2a(i), and "Bill loves x" is shorthand for "Bill has a strong preference for x". Now consider the ratio on the right. If Bill loves x, there is a high probability that he will choose bag 3. However, if Bob loves x, there is perhaps only a medium-high probability that he will choose bag 3 because he might not like e. It follows that the ratio on the right of the expression exceeds one and therefore the ratio on the left exceeds one, hence the stronger attribution for Bill.

So far we have focused on predictions about pairs of decision events, but the model also predicts an ordering over the full space of decision events (see Figure 1b). The model's predicted ordering for four of the events in the experiment is shown in Figure 2a(ii), and this ordering generates predictions about the relative magnitudes of the effects for each pairwise comparison. The arrows are directed from the larger attribution to the smaller attribution, and the thicker and darker the arrow, the larger the predicted difference. Note that these arrows satisfy the property of transitivity: if $D1$ produces a stronger attribution than $D2$, which in turn produces a stronger attribution than $D3$, then the difference in attribution strength for the comparison $(D1, D3)$ should exceed the differences for both $(D1, D2)$ and $(D2, D3)$.

Human judgments Mean human ratings are shown in the second row of Figure 2a(i). The human data were re-centered around 0, meaning that 3.5 is the highest possible rating in the plot. As predicted by the model, the mean ratings are positive in all cases, indicating that the top actor in each pair was at-

tributed a stronger preference for Brand x than the bottom actor. All four of the factors in Figure 1c affected the inferences people made about other people's preferences, and the results for the conflict case (the last bar) suggest that factor F1 has a stronger influence than factor F2. Wilcoxon signed ranked tests were performed for each of the eight pairs. Each bar except the smallest is significantly different from 0 ($p < .02$). The result represented by the smallest bar is marginally significant ($p = .059$).

Figure 2a(iii) shows the ordering diagram derived from the human data. There is no guarantee that the empirical results for all pairs of events will be consistent with a single ordering. The diagram shows, however, that the human data do satisfy the transitivity property, suggesting that a stable psychological ordering of these events exists, and this ordering is consistent with the model's prediction. The model fails to correctly predict the relative differences in strength between adjacent decision events in the ordering. These differences, however, must be interpreted with care. The mean standard deviation across conditions for the human ratings was 1.65, suggesting that the differences may not be reliable.

The effect sizes in Figure 2a(i) differ from those reported by Newtson (1974), but the direction of each effect replicates his results for the six cases that he studied. Newtson, however, was unable to provide a clear explanation for the conflict case involving F1 and F2, and he and others (Jones & McGillis, 1976) have claimed that there is no logical reason for chosen effects to be more informative than forgone effects. As previously discussed, this result is a consequence of our model's basic assumptions. This suggests that probabilistic inference may provide a better account of human inferences and attributions than a strictly logical approach.

Experiment 2

Experiment 1 demonstrated that all four of the factors in Figure 1c shape people's inferences about others' preferences. We also showed that the multinomial logit model accounts for the influence of each of these factors. So far we have focused exclusively on positive effects. However, the model predicts that three of the four factors we considered work in opposite directions when the effects are negative (e.g., electric shocks). In the first two comparisons of Figure 1c(i)–(ii), if Bob chooses to receive two shocks when he could have received one, we might infer that he considers shock x relatively tolerable. In the third comparison (iii), any sensible person would join Alice in choosing one shock over three, but observing Bob's choice suggests that he considers shock x relatively tolerable. Note, however, that the fourth comparison (iv) may lead to the same inference regardless of whether the effects are positive or negative. In each case, Alice chose x three times, suggesting that her preference for x is relatively strong. Our second experiment tested all of these predictions by exploring how the four factors shape inferences about negative effects.

Method

Participants 320 participants were recruited from the Amazon Mechanical Turk website. They were paid for their participation.

Design The experiment consisted of the same eight between-subject conditions that were used in Experiment 1, with 40 participants allocated to each condition. We collected a larger number of participants for Experiment 2 because preliminary results suggested that in some cases an effect was present but relatively small.

Procedure The procedure for Experiment 2 was largely the same as in Experiment 1. The cover story was changed to involve a choice between sets of painful electrical shocks at different body locations so that the effects were unambiguously negative. Accordingly, the question participants were asked was revised to read, "Based only on the above information, which person do you think finds shocks at location x more tolerable?"

Results

Model predictions Model predictions were generated in the same way as for Experiment 1. Because the effects in this experiment were intended to be clearly negative, we used a prior distribution on utilities with mean $\mu = -2\sigma$ that places most of its probability mass is below zero. The predictions for each comparison are shown in the top row of Figure 2b(i). Compared to Experiment 1, the model predicts that the direction for all but two of the comparisons should reverse. This is also reflected in the ordering diagram in Figure 2b(ii).

Human judgments Mean human ratings are shown in the second row of Figure 2b(i). Comparing these results to those from Experiment 1 in Figure 2a(i) suggests that the inferences people draw about actor's preferences when observing choices among negative effects are qualitatively different than for only positive effects. Note that these inferences are not universally reversed, as indicated by the comparison involving factor F4 (number of decisions). However, some of these effects are small. Wilcoxon signed rank tests were performed for each set of human data. The effects depicted by black bars in the figure were significantly different from 0 ($p < .05$). Of the remaining effects, three out of four were in the predicted direction.

These small effects make the ordering diagram in Figure 2b(iii) difficult to interpret. However, of the four cases shown in the diagram, participants' strongest and weakest attributions were for cases predicted by the model.

The mean standard deviation across conditions for the human ratings was 2.09, which is higher than for Experiment 1 (1.65). This may suggest that inferences about negative effects are more difficult than inferences about positive effects, perhaps because choices among negative effects are less familiar in everyday life. An alternative explanation is that electric shocks are unfamiliar examples of negative effects, and that more familiar effects, like doing chores, may be eas-

ier to reason about.

Conclusion

In this paper, we characterized a space of decision events that can be used to explore how people make inferences about other people's preferences. We identified four factors that shape the strength of these inferences: number of chosen effects, number of forgone effects, number of available options, and number of decisions made. In two experiments, we demonstrated that a standard model of choice behavior—the multinomial logit model—predicts the effects of these factors reasonably well.

In particular, the multinomial logit model offers an explanation for why people find the number of chosen effects to be more informative for inferring preferences than the number of forgone effects, an observation that previous theoretical accounts have struggled to explain. It also accounts for the fact that reasoning about negative effects can lead to qualitatively different inferences than for positive effects, rather than a simple reversal of all judgments. This suggests that the model, while fairly generic, may be applicable to a variety of preference learning problems. One question for future work is how people reason about decisions involving choices among options with positive *and* negative effects, which are more like the decisions people make on a daily basis with multiple trade-offs.

Although the factors we examined comprise all the major dimensions of variability in the space of decision events, people's preference attributions are surely influenced by other factors that are independent of the structure of the decision event. For instance, we noted earlier that the model explored in this paper offers a formal account of several principles of the correspondent inference theory of attribution. The theory, however, includes some additional principles that we did not address. One principle concerns the influence of expectations (Jones & Davis, 1965). As an example, suppose Alice is given a choice between a new car and a new bike on a game show. Observing Alice choose the car indicates that she values cars over bikes, but this observation is not likely to change your initial expectations about Alice's preferences, because nearly everyone would consider the car more valuable. By contrast, observing Alice choose the bike would be highly unexpected and this information would likely cause you to drastically alter your beliefs about her preferences.

We did not discuss the difference between expectations and revised beliefs, but this principle is naturally captured using the multinomial logit model by adjusting the priors assigned to different utilities. We might assign a high prior to Alice's utility associated with cars, a low prior to her utility associated with bikes, and an even lower prior to her utility associated with pencils. In other words, although the model in this paper is simple, it is flexible enough to handle many aspects of preference learning and attribution that have interested researchers for some time (Gilbert, 1998).

Acknowledgments We thank Chris Lucas, David Danks, and members of the Danks/Kemp research discussion group for helpful feedback on this research.

References

Bergen, L., Evans, O. R., & Tenenbaum, J. B. (2010). Learning structured preferences. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.*

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365-382.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1). New York: Oxford University Press.

Heider, F. (1958). *The psychology of interpersonal relations*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2). New York: Academic Press.

Jones, E. E., & McGillis, D. (1976). Correspondence inferences and the attribution cube: A comparative reappraisal. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 1). Hillsdale, NJ: Erlbaum.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist*, *28*(2), 107-128.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of others. *Psychological Science*, *21*, 1134-1140.

Lucas, C. G., Griffiths, T. L., Xu, F., & Fawcett, C. (2009). A rational model of preference learning and choice prediction by children. In *Proceedings of Neural Information Processing Systems 21.*

Luce, R. D. (1959). *Individual choice behavior*. John Wiley.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics.* New York: Amademic Press.

Medcof, J. W. (1990). PEAT: An integrative model of attribution processes. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23). New York: Academic Press.

Newtson, D. (1974). Dispositional inference from effects of actions: Effects chosen and effects forgone. *Journal of Experimental Social Psychology*, *10*, 489-496.