Structured, uncertainty-driven exploration in real-world consumer choice

Eric Schulz[a,1,2], **Rahul Bhui**[a,1], **Bradley C. Love**[b,c], **Bastien Brier**[d], **Michael T. Todd**[d], **and Samuel J. Gershman**[a]

[a]Harvard University, Department of Psychology, 52 Oxford St, Cambridge, MA 02138, USA; [b]University College London, Department of Experimental Psychology, 26 Bedford Way, London WC1H 0AP UK; [c]The Alan Turing Institute, 96 Euston Rd, Kings Cross, London NW1 2DB, UK; [d]Deliveroo, Data Science Team, 1 Cousin Lane, London EC4R 3TE, UK

Making good decisions requires people to appropriately explore their available options and generalize what they have learned. While computational models can explain exploratory behavior in constrained laboratory tasks, it is unclear to what extent these models generalize to real world choice problems. We investigate the factors guiding exploratory behavior in a data set consisting of 195,333 customers placing 1,613,967 orders from a large online food delivery service. We find important hallmarks of adaptive exploration and generalization, which we analyze using computational models. In particular, customers seem to engage in uncertainty-directed exploration and use feature-based generalization to guide their exploration. Our results provide evidence that people use sophisticated strategies to explore complex, real-world environments.

Exploration | Generalization | Reinforcement Learning | Decision Making

When facing a vast array of new opportunities, a decision maker has two key tasks: to acquire information (often through direct experience) about available options, and to apply that information to assess options not yet experienced.

These twin problems of *exploration* and *generalization* must be tackled by any organism trying to make good decisions, but they are challenging to solve because optimal solutions are computationally intractable (1). Consequently, the means by which humans succeed in doing so—especially in the complicated world at large—have proven puzzling to psychologists and neuroscientists. Many heuristic solutions have been proposed to reflect exploratory behavior (2–4), inspired by research in machine learning (5, 6). However, most studies have used a small number of options and simple attributes (7). To truly ascertain the limits of exploration and generalization requires empirical analysis of behavior outside the lab.

We study learning and behavior in a complex environment using a large data set of human foraging in the "wild"—online food delivery. Each customer has to decide which restaurant to pick out of hundreds of possibilities. How do they make a selection from this universe of options? Guided by algorithmic perspectives on learning, we look for signatures of adaptive exploration and generalization that have been previously identified in the lab. This allows us not only to characterize these phenomena in a naturally incentivized setting with abundant and multi-faceted stimuli, but also to weigh in on existing debates by testing competing theories of exploratory choice.

We address two broad questions. First, how do people strategically explore new options of uncertain value? Different algorithms have been proposed to describe exactly how uncertainty can guide exploration in qualitatively different ways, such as by injecting *randomness* into choice, or by making choices *directed* toward uncertainty (8). However, results have been mixed, and these phenomena remain to be studied under real-world conditions. Second, how do people generalize their experiences to other options? Modern computational theories make quantitative predictions about how feature-based similarity should govern generalization, which can in turn guide choice. But again it is unclear whether these theories can successfully predict real-world choices.

Our results suggest that customers explore (i.e., order from unexperienced restaurants) adaptively based on signals of restaurant quality, and make better choices over time. Exploration is indeed risky and leads to worse outcomes on average, but people are more likely to explore in cities where this downside is lower due to higher mean restaurant quality. Moreover, we show that customers' exploratory behavior might not only take into account the prospective reward from choosing a restaurant, but also the degree of uncertainty in their reward estimates. Consistent with an optimistic uncertainty-directed exploration policy, they preferentially sample lesser known options and are more likely to reorder from restaurants with higher uncertainties.

Importantly, we apply cognitive and statistical modeling to customers' choice behavior and find that their choices are best fit by a model that includes both an "uncertainty bonus" for unfamiliar restaurants, and a mechanism for generalization by function learning (based on restaurant features). People appear to benefit from such generalization, as exploration yields better realized outcomes in cities where features have more predictive power. We also show that people generalize their experiences across different restaurants within the same broad cuisine type, defined both empirically within the data set, and by independent similarity ratings. As predicted by a combination of similarity-based generalization and uncertainty-directed exploration, good experiences encourage selection of other restaurants within the same category, while bad experiences discourage this to an even greater extent.

Significance Statement

We study how people make choices among a large number of options when they have limited experience. In a large data set of online food delivery purchases, we find evidence for sophisticated exploration strategies predicted by contemporary theories. People actively seek to reduce their uncertainty about restaurants, and employ similarity-based generalization to guide their selections. Our findings suggest that theories of exploratory choice have real-world validity.

ES, RB, and BB extracted and analyzed the data. BCL, MTT and SJG supervised the work. All authors wrote the paper.

No conflict declared.

[1] E.S. and R.B. contributed equally to this work.

[2] To whom correspondence should be addressed: schulz@fas.harvard.edu

In order to set the stage for our analyses of purchasing decisions, we first review the algorithmic ideas that have been developed to explain exploration in the laboratory.

Prior work on the exploration-exploitation dilemma

Uncertainty-guided algorithms. Most of what we know about human exploration comes from *multi-armed bandit tasks*, in which an agent repeatedly chooses between several options and receives reward feedback (9, 10). Since the distribution of rewards for each option is unknown at the beginning of the task, an agent is faced with an *exploration-exploitation dilemma* between two types of actions: should she exploit the options she currently knows will produce high rewards while possibly ignoring even better options? Or should she explore lesser-known options to gain more knowledge but possibly forego high immediate rewards? Optimal solutions only exist for simple versions of this problem (1). These solutions are in practice difficult to compute even for moderately large problems. Various heuristic solutions have been proposed. Generally, these heuristics coalesce around two algorithmic ideas (8). The first one is that exploration happens randomly, for example by occasionally sampling one of the options not considered to be the best (11); or by so-called soft-maximization of the expected utilities for each option—i.e., randomly sampling each option proportionally to its value. The other idea is that exploration happens in a directed fashion, whereby an agent is explicitly biased to sample more uncertain options. This uncertainty-guidance is frequently formalized as an "uncertainty bonus" (5) which inflates an option's expected reward by its uncertainty.

There has been a considerable debate about whether or not directed exploration is required to explain human behavior (12). For example, Daw and colleagues (12) have shown that a softmax strategy explains participants' choices best in a simple multi-armed bandit task. However, several studies have produced evidence for a direct exploration bonus (4, 13). Recent studies have proposed that people engage in both random and directed exploration (2, 14). It has also been argued that directed exploration might play a prominent role in more structured decision problems (15). However, evidence for such algorithms is still missing in real-world purchasing decisions, where other mechanisms such as coherency maximization have been observed (7, 16).

Generalization. Multiple studies have emphasized the importance of generalization in exploratory choice. People are known to leverage latent structures such as hierarchical rules (17) or similarities between a bandit's arms (18).

Inspired by insights from the animal literature (19), Gershman et. al (20) investigated how generalization affects the exploration of novel options using a task in which the rewards for multiple options were drawn from a common distribution. Sometimes this common distribution was "poor" (options tended to be non-rewarding), whereas sometimes the common distribution was "rich" (options tended to be rewarding). Participants sampled novel options more frequently in rich environments than in poor environments, consistent with a form of adaptive generalization across options.

Schulz et al. (21) investigated how contextual information (an option's features) can aid generalization and exploration in tasks where the context is linked to an option's quality by an underlying function. Participants used a combination of functional generalization and directed exploration to learn the underlying mapping from context to reward (see also (22)).

Results

We looked for signatures of uncertainty-guided exploration and generalization in a data set of purchasing decisions from the online food delivery service *Deliveroo* (see Materials and Methods for more details), using both statistical and cognitive modeling. Further analyses and details can be found in the SI Appendix. In the first two sections of the Results, we provide some descriptive characterizations of the data set. In particular, we show that customers learn from past experience and adapt their exploratory behavior over time. Moreover, exploration is systematically influenced by restaurant features and hence amenable to quantification. We then turn to tests of our model-based hypotheses. We find that customers' exploratory behavior can be clustered meaningfully, exhibits several signatures of intelligent exploration which have previously been studied in the lab, and can be captured by a model that generalizes over restaurant features while simultaneously engaging in directed exploration.

Learning and exploration over time. We first assessed if customers learned from past experiences, as reflected in their order ratings over time (Fig. 1a). The order rating is defined as customers' evaluation on a scale between 1 (poor) and 5 (great). Customers picked restaurants they liked better over time: there was a positive correlation between the number of a customer's past orders and her ratings ($r = 0.073$; 99.9% CI: 0.070, 0.076, see SI for further analyses).

Next, we assessed exploratory behavior by creating a variable indicating whether a given order was the first time a customer had ordered from that particular restaurant—i.e., a signature of pure exploration (20). Figure 1b shows the averaged probability of sampling a new restaurant over time (how many orders a customer had placed previously).

Customers sampled fewer new restaurants over time, leading to a negative overall correlation between the number of past orders and the probability of sampling a new restaurant ($r = -0.139$; 99.9% CI: -0.142, -0.136). Exploration also comes at a cost (Fig. 1c), such that explored restaurants showed a lower average rating (mean rating=4.257, 99.9% CI: 4.250, 4.265) than known restaurants (mean rating=4.518, 99.9% CI: 4.514, 4.522).

Customers learned from the outcomes of past orders. Figure 1d shows their probability of reordering from a restaurant as a function of their reward prediction error (RPE; the difference between the expected quality of a restaurant, as measured by the restaurant's average rating at the time of the order, and the actual pleasure customers perceived after they had consumed the order, as indicated by their own rating of the order). RPEs are a key component of theories of reinforcement learning (23), and we therefore expected that customers would update their sampling behavior after receiving either a positive or a negative RPE. Confirming this hypothesis, customers were more likely to reorder from a restaurant after an experience that was better than expected (positive RPE: p(reorder)=0.518, 99.9%; CI: 0.515, 0.520) than after an experience that was worse than expected (negative RPE: p(reorder)=0.394, 99.9%; CI: 0.391, 0.398). The average cor-

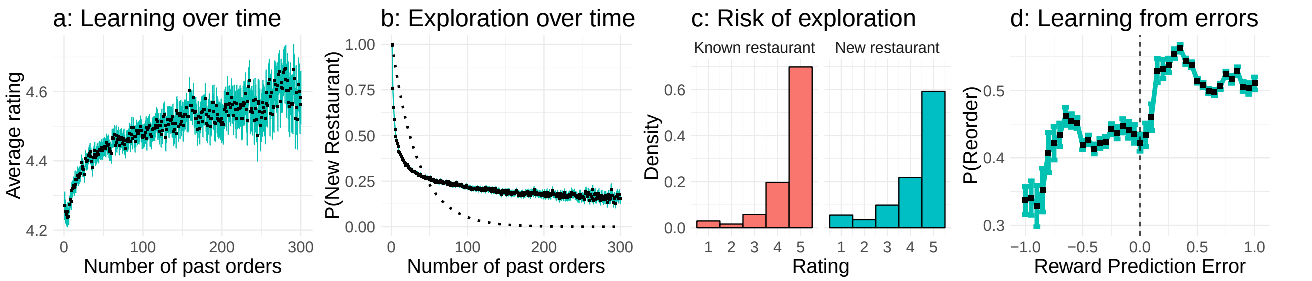

Fig. 1. Learning and exploration over time. a: Average order rating by number of past orders. **b:** Probability of sampling a new restaurant in dependency of the number of past orders. Dashed black line indicates simulated exploratory behavior of agents randomly exploring available restaurants. **c:** Distribution of order ratings for newly sampled and known restaurants. **d:** Average probability of reordering from a restaurant as a function of reward prediction error. Means are displayed as black squares and error bars show the 95% confidence interval of the mean.

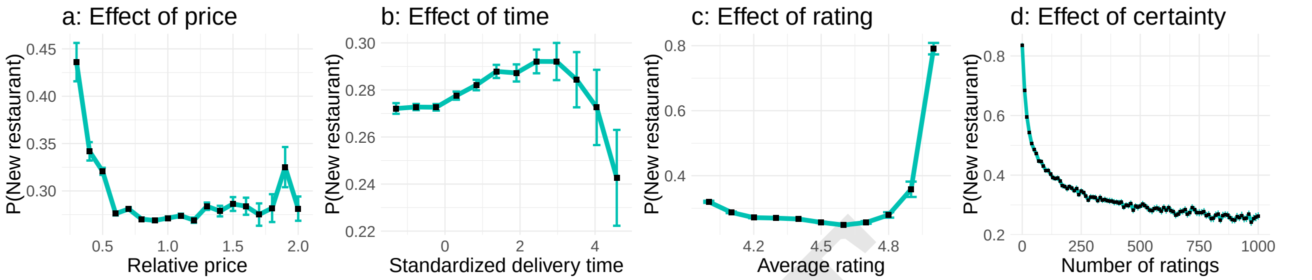

Fig. 2. Factors influencing exploration.
a: Effect of relative price. The relative price indicates how much cheaper or more expensive a restaurant was compared to an average restaurant in the same city. **b:** Effect of standardized (z-transformed) estimated delivery time. **c:** Effect of average rating. **d:** Effect of a restaurant's number of past ratings (certainty). Means are displayed as black squares and error bars show the 95% confidence interval of the mean.

relation between RPEs and the probability of reordering was $r = 0.110$ (99.9% CI: 0.107, 0.114).

Determinants of exploration. In the next part of our analysis, we focused on what factors were associated with the decision to explore a new restaurant. In particular, we assessed if exploratory behavior was systematic and therefore looked at the following four restaurant features that were always visible to customers at the time of their order: the relative price (i.e., how much cheaper or more expensive a restaurant is compared to the average within the same country) of a restaurant, its standardized estimated delivery time, the mean rating of a restaurant at the time of the order, and the number of people who had rated the restaurant before.

Customers preferred restaurants that were comparatively cheaper (Fig. 2a): the correlation between relative price and the probability of exploration was negative ($r = -0.059$; 99.9% CI: -0.0641, -0.0548). There was a non-linear relationship between a restaurant's estimated delivery time and its probability of being explored (Fig. 2b): exploration was most likely for standardized delivery times between 1 and 2.5 (0.288, 99.9% CI: 0.285, 0.292), and less likely for delivery times below 1 (0.288, 99.9% CI: 0.285, 0.292 or above 2.5 (0.252, 99.9% CI: 0.229, 0.274). This indicates that customers might have taken into account how long it would take to plausibly prepare and deliver a good meal when deciding which restaurants to explore. The average rating of a restaurant also affected customers' exploratory behavior (Fig. 2c): higher ratings were associated with a higher chance of exploration ($r = 0.038$; 99.9% CI: 0.0337, 0.0430). The number of ratings per restaurant also influenced exploration (Fig. 2d), with a negative correlation of $r = -0.188$ (99.9% CI: -0.192, -0.183). This may have a mechanical component because restaurants that have been tried more frequently are intrinsically less likely to be explored for the first time. We therefore repeated this analysis for all restaurants that had been rated more than 500 times, yielding a correlation of $r = -0.034$ (99.9% CI: -0.042, -0.026).

Table 1. Results of the mixed-effects logistic regression.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.663	0.008	-82.01	<.001
Relative price	-0.014	0.006	-2.27	.02
Time-Linear	-0.0246	0.008	-3.22	.001
Time-Quadratic	0.015	0.004	3.89	<.001
Average rating	0.086	0.006	13.85	<.001
Number of ratings	-0.475	0.007	-70.27	<.001

We standardized and entered all of the variables into a mixed-effects logistic regression modeling the exploration variable as the dependent variable and adding a random intercept for each customer (see SI for full model comparison). We again found that a smaller number of total ratings ($\beta = -0.475$), a higher average rating ($\beta = 0.086$), and a lower price ($\beta = -0.014$) as well as a quadratic effect of time ($\beta_{\text{Linear}} = -0.025$, $\beta_{\text{Quadratic}} = 0.015$) were all predictive of customers' exploratory behavior.

In summary, exploration in the domain of online ordering is systematic, interpretable and amenable to quantification. We next turned to an examination of our model-based hypotheses concerning directed exploration and generalization.

Signatures of uncertainty-directed exploration

We probed the data for signatures of uncertainty-directed exploration algorithms that attach an uncertainty bonus to each option. One such signature is that directed and random exploration make diverging predictions about behavioral changes after either a positive or a negative outcome. Whereas random (softmax) exploration predicts no difference between the extent of sampling behavior change following a better-than-expected outcome versus following a worse-than-expected outcome, directed exploration predicts a stronger increase in sampling behavior after a worse-than-expected outcome (see SI). This is

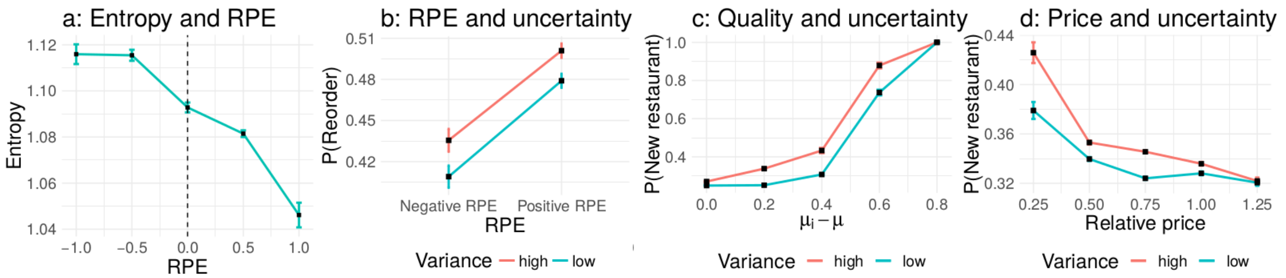

Fig. 3. Signatures of uncertainty-directed exploration.

a: Entropy of the next 4 choices in dependency of reward prediction error (RPE). **b:** Probability of reordering from a restaurant in dependency of RPE, shown for restaurants with high and low relative variance. **c:** Probability of choosing a novel restaurant in dependency of its difference to an average restaurant within the same cuisine type for restaurants with high and low relative variance. **d:** Probability of choosing a novel restaurant in dependency of its relative price for restaurants with high and low relative variance.

due to the properties of algorithms that assess an option's utility by a weighted sum of its expected reward and its standard deviation. After a bad experience, the mean and standard deviation both go down, whereas after a good experience the mean goes up but the standard deviation goes down. Thus, there should be greater change in customers' sampling behavior after a bad than after a good outcome.

We verified this prediction by calculating the Shannon entropy of customers' next 4 purchases after having experienced either a better-than or a worse-than-expected order. The calculated entropy was higher for negative RPEs (Fig 3a; 1.112, 99.9% CI: 1.109, 1.115) than for positive RPEs (1.082, 99.9% CI: 1.081, 1.084), in line with theoretical predictions of a directed exploration algorithm.

We calculated each restaurant's relative variance, i.e., how much more variance in its ratings a restaurant possessed as compared to the average variance per restaurant within the same cuisine type (although customers cannot see the actual estimate of a restaurant's variance in ratings, they can access all past rating as well as a summary that shows the distribution over ratings). We then compared the reorder probability for restaurants with a high vs. low relative rating variance, based on a median split (Fig. 3b). This probability was higher for restaurants with high relative variance than for restaurants with low relative variance for both negative and positive RPEs. Thus, customers were more likely to return to restaurants with higher relative uncertainty.

We also assessed customers' exploratory behavior in dependency of the differences in ratings for a given restaurant as compared to the average of all restaurants within the same cuisine type (value difference). The probability of exploring a new restaurant increased as a function of the restaurant's value difference (Fig. 3c; $r = 0.05$, 99.9% CI: 0.045, 0.056). Additionally, a restaurant's relative variance also correlated with its probability of being explored (Fig. 3c; $r = 0.05$; 99.9% CI: 0.045, 0.056). Comparing restaurants with a high vs. low relative variance in their ratings revealed a shift of the choice function towards the left. In other words, restaurants with higher relative uncertainty (0.344; 99.9% CI: 0.341, 0.349) are preferred to restaurants with lower relative uncertainty (0.319; 99.9% CI: 0.317, 0.321), as predicted by uncertainty-directed exploration strategies (2). This difference can also be observed when repeating the same analysis using a restaurant's price (Fig. 3d): as restaurants get more expensive, they are less likely to be explored ($r = -0.017$; 99.9%CI: -0.023, -0.013). This function is again shifted for restaurants with higher relative uncertainty: given a similar price range, relatively more uncertain restaurants are more likely to be explored than less uncertain restaurants.

Table 2. Results of mixed-effects logistic regression.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.342	0.007	45.81	<.001
Value difference	0.114	0.0135	8.47	<.001
Relative price	-0.087	0.007	-11.67	<.001
Variance difference	0.084	0.003	24.13	<.001

To further validate these findings, we fit a mixed-effects logistic regression, using the exploration variable as the dependent variable. For the independent variables, we used the mean difference in ratings between the restaurant and the average restaurant within the same cuisine type, a restaurant's relative price, and its relative uncertainty (see Tab. 2). The average value difference ($\beta = 0.114$), the relative price $\beta = -0.0876$) and the relative uncertainty ($\beta = 0.084$) all affected a restaurants' probability to be explored. Thus, even when taking into account a restaurant's price and its ratings, customers still preferred more uncertain options. This provides further evidence for a directed exploration strategy.

Signatures of generalization. Having observed how exploratory behavior changes with experience, we investigated how generalization might affect exploration in several ways. First, we looked for evidence of information spillovers by analyzing changes in exploration within cuisine clusters. These seven clusters were defined in a data-driven manner based on patterns of consecutive explorations, that is, how one exploratory choice predicted the next (see Fig. 4a and Materials and Methods). This was also related to a subjective understanding of similarity; the frequency of switching between cuisine types was strongly correlated with similarity ratings provided by 200 workers on Amazon Mechanical Turk ($r = 0.78$; Fig. 5a). Hinting at strategies of directed exploration as before, we found that bad outcomes had a larger effect than good outcomes compared to a baseline of average switches (Fig. 4b)—customers were especially averse to exploring other restaurants in the same cluster after a worse-than-expected outcome (-5.19%), more than they favored such exploration after a better-than-expected outcome (+2.27%). This suggests that uncertainty-modulated exploration takes into account experiences with different restaurants of similar types. Intriguingly, we also observed that customers tended to switch to exploring "Unhealthy" cuisines after bad experiences with any other type (+2.72%). This may reflect people balancing differing goals across successive choices (24).

Second, we analyzed how exploration is modulated by the distribution of restaurant quality in a city. Gershman et al. (20) showed that participants explore novel options more

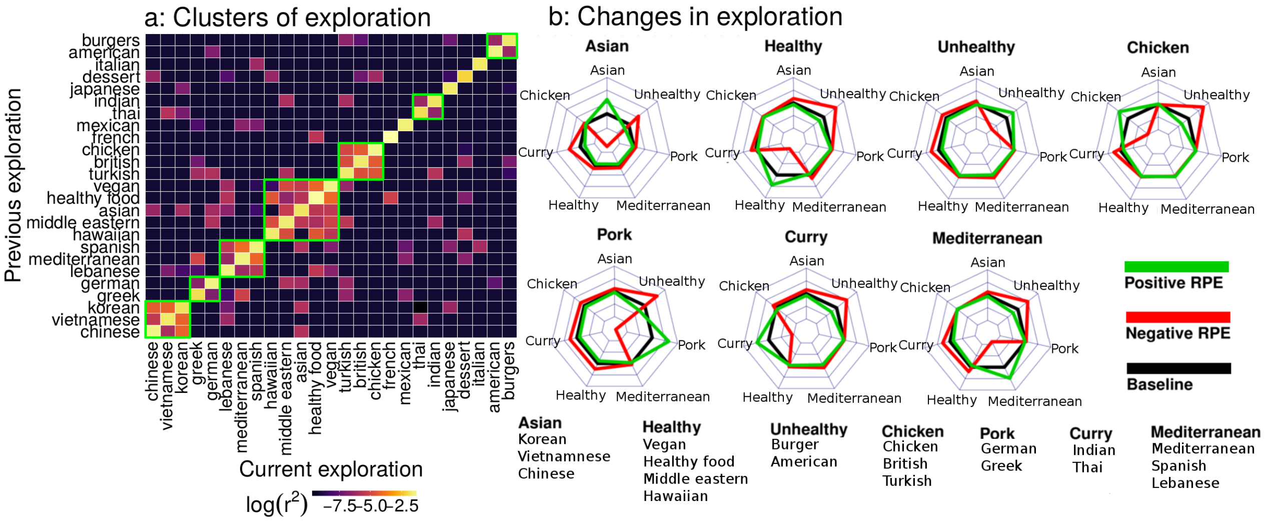

Fig. 4. Clusters and changes of exploration.

a: Clusters of exploration between different cuisine types within customers' consecutive explorations. Green rectangles mark clusters of exploration. **b:** Moves between clusters after better-than-expected (positive RPE) and worse-than-expected (negative RPE) outcomes as compared to a restaurant-specific mean baseline. Centers of radar plots indicate a change of -5%, outermost lines indicate a change of +5%. A change of 1% roughly translates to 500 orders.

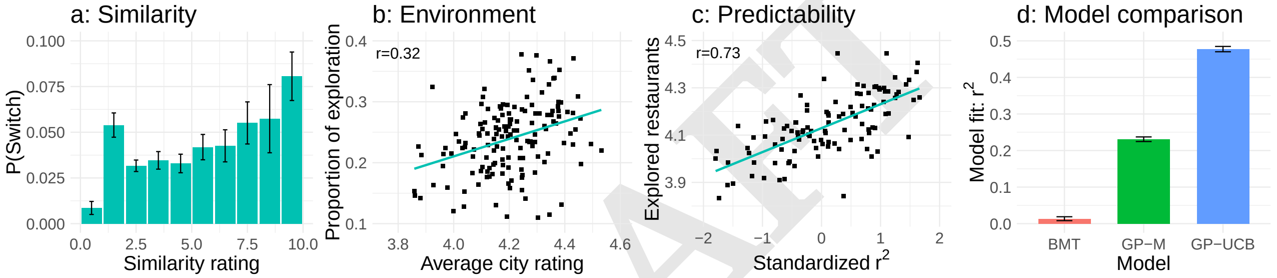

Fig. 5. Signatures of generalization.

a: Probability of switches between cuisine types and rated similarities between the same types. **b:** Average rating per city and proportion of exploratory choices. Turquoise line marks least-square regression line. **c:** Predictability of a restaurant's quality and average rating of explored restaurants. Turquoise line marks least-square regression line. **d:** Results of model comparison for new customers' behavior. Considered models were the Bayesian Mean Tracker (BMT), a Gaussian Process with a mean-greedy sampling strategy (GP-M), and a Gaussian Process with a Upper Confidence Bound sampling strategy (GP-UCB).

frequently in environments where all options are generally good. We found evidence for this phenomenon in our data (Fig. 5b): there was a positive correlation between a city's average restaurant rating and the proportion of exploratory choices in that city ($r = 0.32$; 99.9% CI: 0.21, 0.49, see SI Appendix for partial correlations). Moreover, there was also a positive correlation between a city's variance of ratings and the proportion of exploratory choices ($r = 0.48$; 99.9% CI: 0.37, 0.59), indicating that higher uncertainties in ratings were linked to more exploration.

Third, we examined how the success of exploration depended on the predictability of individual ratings from restaurant features (price, delivery time, mean rating, and number of ratings). Customers gave higher ratings to explored restaurants in cities where ratings were generally more predictable ($r = 0.73$; Fig. 5c, 99.9% CI: 0.53, 0.84). Thus, exploration seemed to be enhanced by the degree to which features permitted a reduction in uncertainty, similar to findings in contextual bandit tasks (21).

In an attempt to test algorithms of both directed exploration and generalization simultaneously, we compared three models of learning and decision making based on how well they captured the sequential choices of 3,772 new customers who had just started ordering food and who had rated all of their orders. The first model was a Bayesian Mean Tracker (BMT) that estimates the mean quality for each restaurant independently. The second model was an extension of the BMT model (Gaussian Process regression) that estimates mean quality as a function of observable features (price, mean rating, delivery time, and number of past ratings). The shared feature space allows this model to generalize across restaurants. Gaussian Process regression is a powerful model of generalization that has been applied to model how participants learn latent functions to guide their exploration (15, 21, 22). It can be seen as a Bayesian variant of similarity-based decision making, akin to economic theories of case-based decision making (25) and psychological formulations of similarity judgments (26). This model was paired with two different policies: stochastic sampling of actions in proportion to their estimated mean quality (GP-M), or with a directed exploration strategy that sampled based on both the mean and an uncertainty bonus (formally, an option's upper confidence bound, GP-UCB). We treated customers' choices as the arms of a bandit and their order ratings as their utility, and then evaluated each model's performance based on its one-step-ahead prediction error, standardizing performance by comparing to a random baseline. Since it was not possible to observe all restaurants a customer might have considered at the time of an order, we compared the different models based on how much higher in utility they predicted a customer's final choice compared to an option with average features out of all the restaurants available in that customer's city. As Fig. 5d shows, the BMT model barely performed above chance ($r^2 = 0.013$; 99.9% CI: 0.005, 0.022). Although the GP-M model performed better than the BMT model ($r^2 = 0.231$; 99.9% CI: 0.220, 0.241), the GP-UCB model achieved by far the best performance ($r^2 = 0.477$; 99.9% CI: 0.465, 0.477). Thus, a sufficiently predictive model of customers' choices required both a mechanism of generalization (learning how features map onto rewards), and a directed exploration strategy (combining a restaurant's mean and uncertainty to estimate its decision value).

Discussion

We investigated customers' exploratory behavior in a large data set of online food delivery purchases. Customers learned from past experiences, and their exploration was affected by a restaurant's price, average rating, number of ratings and estimated delivery time. Our results further provide evidence for several theoretical predictions: people engaged in uncertainty-directed exploration, and their exploration was guided by similarity-based generalization. Computational modeling showed that these patterns could be captured quantitatively.

Of course, drawing causal inferences from large data sets is difficult (27). Thus, although we believe that our results provide evidence that people use sophisticated strategies in complex, naturalistic environments, these effects nonetheless deserve further investigation, for example by conducting online experiments.

Furthermore, our model does currently not explain all possible intentions customers might have when ordering food such as maintaining a healthy diet or balancing different goals over successive choices like saving money and trying out expensive food (24). These could hypothetically be incorporated into the kernel function.

Taken together, our results advance our understanding of human choice behavior in complex real-world environments. The results may also have broader implications for understanding consumer behavior. For example, we found that customers frequently switch to unhealthy food options after bad experiences. A potential strategy to increase the exploration of healthy food might thus be to increase healthy restaurants' relative uncertainty by grouping them with other frequently explored options such as Asian restaurants, which showed a comparatively lower relative uncertainty per restaurant.

While we have focused on using cognitive models to predict human choice behavior, the same issues come up for the design of recommendation engines in machine learning. These engines use sophisticated statistical techniques to make predictions about behavior, but do not typically try to pry open the human mind (28). This is a missed opportunity, since one could generate better recommendations of which restaurants to try next, based on a particular customer's estimated values and uncertainties; as models of human and machine learning have become increasingly intertwined, insights from cognitive science may help build more intelligent machines for predicting and aiding consumer choice.

ACKNOWLEDGMENTS. ES was supported by the Harvard Data Science Initiative. SJG was supported by the Office of Naval Research under Grant N000141712984.

Materials and Methods

The Deliveroo data set. The data consisted of a representative random subset of customers ordering food from the online food delivery service "Deliveroo". The data set contained 195,333 fully anonymized customers. These customers placed 1,613,968 orders over two month (February and March 2018) in 197 cities. There were 30,552 restaurants in total leading to an average of 155 restaurants per city. We arrived at this data set by filtering out customers with less than 5 orders (too little data points to analyze learning) and more than 100 orders (likely multiple people sharing an account).

Clustering analysis. Cuisine tags were manually defined by Deliveroo. We analyzed for each cuisine type how much exploring this type on a time point t was predictive of exploring another cuisine type on a time point $t + 1$, using a linear regression model. Repeating this analysis for every combination of cuisine types lead to the graph shown in Figure 4a. We then analyzed the resulting matrix of r^2-values using hierarchical clustering. This clustering excluded the cuisine type "European" as it was found to contain little information about customer choice behavior.

Similarity judgments. To elicit similarity ratings between different cuisine types, we asked 200 participants on Amazon's Mechanical Turk to rate the similarities between two randomly sampled types out of the 20 types used for the clustering analysis reported above. Participants were paid \$1 and had to rate 50 pairs of cuisine types on a scale from 0 (not at all similar) to 10 (totally similar).

1. Whittle P (1980) Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 143–149.
2. Gershman SJ (2018) Deconstructing the human algorithms for exploration. *Cognition* 173:34–42.
3. Speekenbrink M, Konstantinidis E (2015) Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science* 7(2):351–367.
4. Frank MJ, Doll BB, Oas-Terpstra J, Moreno F (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience* 12(8):1062.
5. Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3(Nov):397–422.
6. Srinivas N, Krause A, Kakade SM, Seeger MW (2012) Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory* 58(5):3250–3265.
7. Riefer PS, Prior R, Blair N, Pavey G, Love BC (2017) Coherency-maximizing exploration in the supermarket. *Nature Human Behaviour* 1(1):0017.
8. Schulz E, Gershman SJ (2019) The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology* 55:7–14.
9. Cohen JD, McClure SM, Angela JY (2007) Should I stay or should I go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 362(1481):933–942.
10. Mehlhorn K, et al. (2015) Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision* 2(3):191.
11. Sutton RS, Barto AG, Bach F, , et al. (1998) *Reinforcement learning: An introduction.* (MIT press).
12. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441(7095):876.
13. Knox WB, Otto AR, Stone P, Love B (2012) The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology* 2:398.
14. Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD (2014) Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General* 143(6):2074.
15. Wu CM, Schulz E, Speekenbrink M, Nelson JD, Meder B (2018) Generalization guides human exploration in vast decision spaces. *Naure Human Behaviour* 2:915–924.
16. Todd PM (2017) Human behaviour: Shoppers like what they know. *Nature* 541(7637):294.
17. Badre D, Kayser AS, D'Esposito M (2010) Frontal cortex and the discovery of abstract action rules. *Neuron* 66(2):315–326.
18. Wimmer GE, Daw ND, Shohamy D (2012) Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience* 35(7):1092–1104.
19. Noble J, Todd PM, Tucif E (2001) Explaining social learning of food preferences without aversions: an evolutionary simulation model of norway rats. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268(1463):141–149.
20. Gershman SJ, Niv Y (2015) Novelty and inductive generalization in human reinforcement learning. *Topics in Cognitive Science* 7(3):391–415.
21. Schulz E, Konstantinidis E, Speekenbrink M (2018) Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 44(6):927–943.
22. Schulz E, Franklin NT, Gershman SJ (2018) Finding structure in multi-armed bandits. *bioRxiv* p. 432534.
23. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275(5306):1593–1599.
24. Dhar R, Simonson I (1999) Making complementary choices in consumption episodes: Highlighting versus balancing. *Journal of Marketing Research* 36(1):29–44.
25. Bhui R (2018) Case-based decision neuroscience: Economic judgment by similarity in *Goal-Directed Decision Making.* (Elsevier), pp. 67–103.

26. Goldstone RL (1994) Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(1):3.

27. Shiffrin RM (2016) Drawing causal inference from big data. *Proceedings of the National Academy of Sciences* 113(27):7308–7309.

28. Griffiths TL (2015) Manifesto for a new (computational) cognitive revolution. *Cognition* 135:21–23.

29. Peters J, Mooij JM, Janzing D, Schölkopf B (2014) Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research* 15(1):2009–2053.