Supporting Information

Schulz et al. 10./pnas.XXXXXXXXXX

Data set

We used the following variables to analyze customers' exploratory behavior: anonymized customer ID, anonymized restaurant ID, anonymized order ID, the name of the city in which an order was placed, the cuisine type of the order (180 types in total), the standardized price of the restaurant indicating how much more expensive it was than an average restaurant within the same city (from 0.25 to 2), the standardized estimated delivery time of the order (z-scores from -2 to 3), how many orders a customer had placed previously (over her entire order history), whether or not it was the first time a customer had ordered from the chosen restaurant (again over the entire history), the mean rating of the restaurant at the time of the order from 4 to 5*, the number of previous ratings for the restaurant at the time of the order, and the eventual rating the customer provided (from 1 to 5). We also calculated every order's RPE by subtracting the mean restaurant rating from the eventual order rating. Customers ordered 9.4 times on average during the two months in our subset of delivery data, with a mean time between orders of 5.2 days. On the Deliveroo website, customers saw restaurants ordered based on the companies recommendation system, which marked "old favourites" but also recommended other restaurants similar to the ones a user had previously ordered from and enjoyed. No customer ever ran out of new restaurants to explore in our data set. Figure S1 shows how the distribution of past ratings is presented to customers on the Deliveroo website. The distributions of all variables are shown in Figure S2.

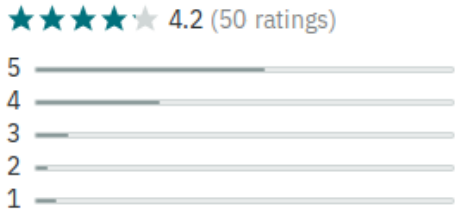

Fig. S1. Screenshot of how the distribution of past ratings is presented to customers on the Deliveroo website. Customers could access the distribution of ratings by clicking on "Show rating details".

Model comparison

New customers data set. For the model comparison, we created a data set containing only customers who had just started ordering food on the Deliveroo website (i.e., new customers). Moreover, we filtered out all customers who did not rate all of their orders. This resulted in a data set of 3,772 customers in total. We used this data set to compare different models of learning combined with different decision strategies, treating customers' chosen restaurants as the arm of a bandit and their ratings as the resulting reward.

Gaussian Process. We use Gaussian Process (GP) regression as a Bayesian model of generalization. A GP is defined as a collection of points, any subset of which is multivariate Gaussian. Let $f : \mathcal{X} \to \mathbb{R}^n$ denote a function over input space $\mathcal{X}$ that maps to real-valued scalar outputs. This function can be modeled as a random draw from a GP:

$$f \sim \mathcal{GP}(m, k), \quad [1]$$

where m is a mean function specifying the expected output of the function given input $\mathbf{x}$, and k is a kernel function specifying the covariance between outputs:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \quad [2]$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right] \quad [3]$$

We fix the prior mean to the mean value of ratings within a given city and use the kernel function to model generalization over the restaurant-specific features.

Conditional on observed data $\mathcal{D}_t = \{\mathbf{x}_j, y_j\}_{j=1}^t$, where $y_j \sim \mathcal{N}(f(\mathbf{x}_j), \sigma_j^2)$ is drawn from the underlying function with added noise $\sigma_j^2 = 1$, we can calculate the posterior predictive distribution for a new input $\mathbf{x}_*$ as a Gaussian:

$$\mathbb{E}[f(\mathbf{x}_*)|\mathcal{D}_t] = m_t(\mathbf{x}_*) = \mathbf{k}_*^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}_t \quad [4]$$

$$\mathbb{V}[f(\mathbf{x}_*)|\mathcal{D}_t] = v_t(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_\star^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_*, \quad [5]$$

where $\mathbf{y} = [y_1, \ldots, y_t]^\top$, $\mathbf{K}$ is the $t \times t$ covariance matrix evaluated at each pair of observed inputs, and $\mathbf{k}_* = [k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_t, \mathbf{x}_*)]$ is the covariance between each observed input and the new input $\mathbf{x}_*$.

To model customers' generalization over restaurants' features, we assume that customers can use the presented features at the time of an order to predict a restaurant's quality, i.e., how much they will like it. These features are the price, the mean rating, the number of past ratings, and the delivery time.

Radial Basis Function kernel. We use a Radial Basis Function (RBF) kernel as a component of the $\mathcal{GP}$ algorithm of generalization. The RBF kernel specifies the correlation between inputs $\mathbf{x}$ and $\mathbf{x}'$ as

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{\lambda}\right), \quad [6]$$

where λ is a length-scale parameter controlling function smoothness. This kernel defines a universal function learning engine based on the principles of Bayesian regression and can model any stationary function.

Bayesian Mean Tracker. The Bayesian Mean Tracker model is implemented as a Bayesian updating model, which assumes no temporal dynamics. In contrast to the GP regression model (which also assumes constant means over time), the Mean Tracker learns the rewards of each restaurant independently, by computing an independent posterior distribution for the

*In total, 94% of the restaurants had higher ratings than 4 and behavior for restaurants with an average rating lower than 4 was unstable. None of the main results change when analyzing the full data set, but estimates for this part of the space were unreliable.

www.pnas.org/cgi/doi/10.1073/pnas.1821028116

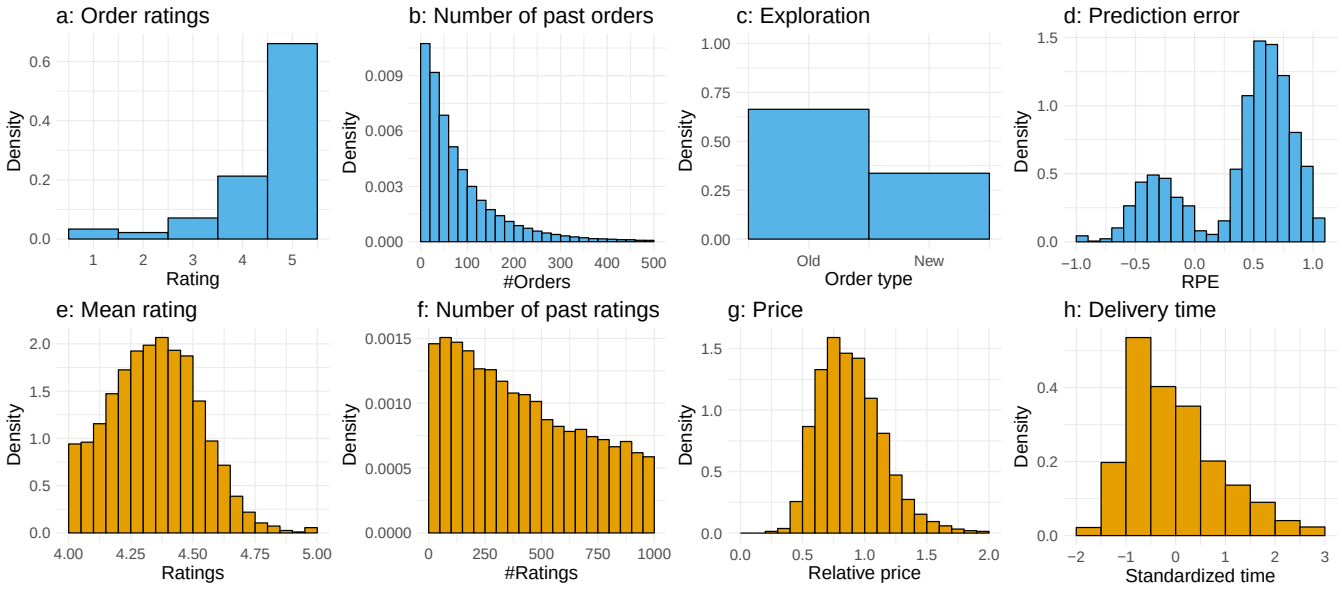

Fig. S2. Distributions of customer-specific (blue) and restaurant-specific (orange) variables.
a: Customers' order ratings from 1 (low quality) to 5 (high quality). **b:** Number of past orders per customer. **c:** Proportion of choosing an old vs. a new restaurant. **d:** Prediction error, defined as the difference between the actual order rating and the mean restaurant rating. **e:** Mean restaurant ratings at the time of an order. Ratings have been truncated at a value of 4 to avoid instability induced by infrequent low average ratings. **f:** Number of past ratings per restaurant at the time of an order. **g:** Relative price per restaurant at the time of an order. **h:** Standardized delivery time at the time of an order.

mean μ_j for each restaurant j. We implemented a version that assumes rewards are normally distributed (as in the GP model), with a known variance but unknown mean, where the prior distribution of the mean is a normal distribution. This implies that the posterior distribution for each mean is also a normal distribution:

$$p(\mu_{j,t}|\mathcal{D}_{t-1}) = \mathcal{N}(m_{j,t}, v_{j,t}) \quad [7]$$

For a given option j, the posterior mean $m_{j,t}$ and variance $v_{j,t}$ are only updated when it has been selected at trial t:

$$m_{j,t} = m_{j,t-1} + \delta_{j,t} G_{j,t} \left[y_t - m_{j,t-1}\right] \quad [8]$$
$$v_{j,t} = \left[1 - \delta_{j,t} G_{j,t}\right] v_{j,t-1} \quad [9]$$

where $\delta_{j,t} = 1$ if option j is chosen on trial t, and 0 otherwise. Additionally, y_t is the observed reward at trial t, and $G_{j,t}$ is defined as:

$$G_{j,t} = \frac{v_{j,t-1}}{v_{j,t-1} + \theta_\epsilon^2} \quad [10]$$

where θ_ϵ^2 is the error variance, which we fixed to 0.2. Intuitively, the estimated mean of the chosen option $m_{j,t}$ is updated based on the difference between the observed value y_t and the prior expected mean $m_{j,t-1}$, multiplied by $G_{j,t}$. At the same time, the estimated variance $v_{j,t}$ is reduced by a factor of $1 - G_{j,t}$, which is in the range $[0, 1]$. The error variance (θ_ϵ^2) acts as an inverse sensitivity, where smaller values result in more substantial updates to the mean $m_{j,t}$, and larger reductions of uncertainty $v_{j,t}$. We set the prior mean to the mean value of all restaurants within a city and the prior variance to $v_{j,0} = 5$.

This model does not generalize at all and can therefore only learn about a restaurant's quality by sampling it. Thus, it predicts that every novel restaurant will just be as good as the average of all restaurants in a city.

Sampling strategies.

Given the normally distributed posteriors of the expected rewards, which have mean $\mu(\mathbf{x})$ and uncertainty (formalized here as standard deviation) $\sigma(\mathbf{x})$, for each restaurant $\mathbf{x}$ (for the Mean Tracker, we let $\mu(\mathbf{x}) = m_{j,t}$ and $\sigma(\mathbf{x}) = \sqrt{v_{j,t}}$, where j is the index of the restaurant characterized by $\mathbf{x}$), we assess different sampling strategies that make probabilistic predictions about how much customers will like a given restaurant. In particular, we combine the Bayesian Mean Tracker with a mean-greedy sampling strategy and the Gaussian Process regression with both a mean-greedy and an upper confidence bound sampling strategy (details below).

Upper Confidence Bound sampling. Given the posterior predictive mean $\mu(\mathbf{x})$ and its attached standard deviation $\sigma(\mathbf{x}) = \sqrt{\sigma^2(\mathbf{x})}$, we calculate the upper confidence bound using a weighted sum

$$\mathrm{UCB}(\mathbf{x}) = \mu(\mathbf{x}) + \beta\sigma(\mathbf{x}), \quad [11]$$

where the exploration factor β determines how much reduction of uncertainty is valued (relative to exploiting known high-value options). We fix $\beta = 1$ for our model comparison, indicating a tendency towards directed exploration.

Mean Greedy Exploitation. A special case of the Upper Confidence Bound sampling strategy (with $\beta = 0$) is a greedy exploitation component that only evaluates points based on their expected rewards

$$\mathrm{M}(\mathbf{x}) = \mu(\mathbf{x}), \quad [12]$$

This sampling strategy only samples options with high expected rewards, i.e., greedily exploits the environment.

Model comparison

We fit all models to a customers' data until time point t and then make predictions about choices on time point $t + 1$.

We apply a softmax choice rule to transform each model's prediction into a probability distribution over options:

$$p(\mathbf{x}) = \frac{\exp(q(\mathbf{x}))}{\sum_{j=1}^{N} \exp(q(\mathbf{x}_j))}, \quad [13]$$

where $q(\mathbf{x})$ is the predicted value of each option $\mathbf{x}$ for a given model (e.g., $q(\mathbf{x}) = \mathrm{UCB}(\mathbf{x})$ for the UCB model).

One-step ahead prediction errors. We fit all models—per customer—to the data a customer has seen until time point t and then make forecasts about choices at time point $t+1$. For example, the Gaussian Process model is fitted to all past restaurants a customer has sampled, using the restaurant's features (i.e., prize, mean rating, number of ratings and delivery time) as the independent variables and the customer's ratings as the dependent variable. Afterwards, it can be used to make predictions about other restaurants' expected ratings (and uncertainties), that can be mapped onto probabilities. The difference between the Gaussian Process model and the Bayesian Mean Tracker is that the Bayesian Mean Tracker does not use any generalization over features, but only updates its predictions (which are equated to the overall mean at the beginning) by sampling a restaurant. The difference between the mean-greedy GP-M model and the GP-UCB model containing a directed exploration component is that the GP-M model equates a restaurant's utility with the predicted mean rating, whereas the GP-UCB model equates a restaurant's utility with its upper confidence bound.

Crucially, it is never possible to assess all restaurants a customer looked at and could have ordered from at a particular time point. We therefore compare the utility of the chosen restaurant to an average restaurant in the same city. For example, for the Gaussian Process model, we compared how much more likely a customer's choice was compared to a restaurant with average feature values. For the BMT model, we compare the assessed utility to the overall average of restaurants in a city.

Predictive accuracy. The error of predictions (computed as predictive log loss) is summed up over all one-step ahead predictions, and is reported as *predictive accuracy*, using a pseudo-R^2 measure that compares the total log loss for each model to that of a random model:

$$R^2 = 1 - \frac{\log \mathcal{L}(\mathcal{M}_k)}{\log \mathcal{L}(\mathcal{M}_{\mathrm{rand}})}, \quad [14]$$

where $\log \mathcal{L}(\mathcal{M}_{\mathrm{rand}})$ is the log loss of a random model (i.e., picking options with equal probability) and $\log \mathcal{L}(\mathcal{M}_k)$ is the log loss of model k's one-step-ahead prediction error. A $R^2 = 0$ corresponds to a prediction accuracy equivalent to chance, while $R^2 = 1$ corresponds to a theoretically perfect predictive accuracy, since $\log \mathcal{L}(\mathcal{M}_k)/\log \mathcal{L}(\mathcal{M}_{\mathrm{rand}}) \to 0$ when $\log \mathcal{L}(\mathcal{M}_k) \ll \log \mathcal{L}(\mathcal{M}_{\mathrm{rand}})$.

Statistical tests

As our data set was large, almost any comparison would be significant at the $\alpha = 0.05$-level. We therefore report the means and 99.9% confidence intervals for each group when reporting differences. We believe that this descriptive comparison makes the size of the differences more interpretable.

Mixed-effects regression

We report the step-wise results for both mixed-effects regression analyses. We compare models based on their Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Factors influencing exploration. For the first mixed-effects regression, we regressed a restaurant's price (Price), mean rating (Rating), number of past ratings (#Ratings), and estimated delivery time (Time) onto whether or not a customer explored that restaurant. Additionally, we entered a random intercept for each customer.

Table S1. Results of mixed-effects logistic regression analyzing determinants of exploration.

Model	AIC	BIC	χ^2	Pr(>χ^2)
Intercept only	258107	258127	–	–
Price	258064	258095	45	<.001
Price+Rating	257294	257334	772	<.001
Price+Rating+#Ratings	251784	251835	5511	<.001
Price+Rating+#Ratings+Time^2	251772	251843	16	<.001

The variables price, rating and the number of ratings all had a linear effect onto a restaurant's probability of being explored, whereas the average delivery time had a nonlinear effect (the expression Time^2 in Tab. S1 indicates that we entered both a linear and a quadratic effect of time into the final model). The final model contained all variables and had a fit of BIC=251772.

Signatures of directed exploration. For the second mixed-effects logistic regression, we regressed a restaurant's value difference (Value), relative uncertainty (Uncertainty) and price (Price) onto the exploration variable. The value difference is defined as the difference in ratings for a given restaurant compared to the average of all restaurants within the same cuisine type. The relative variance is defined as the difference between a restaurant's variance in ratings and the average variance per restaurant within the same cuisine type. The price is the relative price indicating how much more expensive a restaurant was compared to the city's average restaurant price.

Table S2. Results of mixed-effects logistic regression analyzing signatures of directed exploration.

Model	AIC	BIC	χ^2	Pr(>χ^2)
Intercept only	258107	258127	–	–
Value	257184	257214	924	<.001
Value+Price	257152	257193	34	<.001
Value+Price+Uncertainty	257066	257117	88	<.001

The final model contained all three variables and produced a fit of BIC=257117. Thus, relative uncertainty was a significant contributor to customers' exploration behavior beyond value difference and relative price, a strong signature of directed exploration.

Using other measures of uncertainty. We consider another measure of relative uncertainty that takes into account both the distribution and the number of ratings per restaurant. This aggregated measure of uncertainty is the standard error of a restaurant's rating, i.e., $\sigma/\sqrt{n}$. The results (see Tab. S3) showed the same pattern as before with higher values, lower prices and –importantly– higher standard errors leading to increased exploration.

Table S3. Results of mixed-effects logistic regression.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.328	0.007	-43.86	<.001
Value difference	0.147	0.0133	11.09	<.001
Relative price	-0.102	0.008	-13.57	<.001
Standard error difference	0.504	0.015	33.86	<.001

Controlling for non-linear effects of value. Another concern when analyzing signatures of directed exploration is that value, i.e., the average rating of a restaurant, might have a non-linear effect onto customers' probability of exploring a restaurant. For example, if customers give extra weight to high ratings, then this could also link higher variances to exploration behavior. To rule out this possibility, we asses a model that relates restaurant ratings to choices non-linearly by using polynomial (i.e., quadratic and cubic) regression and additionally entering a restaurant's price and uncertainty to this model. Doing so, we find that the positive effect of uncertainty remains unchanged for both cases, when including a quadratic relation between ratings and exploration (estimates for uncertainty: $\beta = 0.083$, $z = 24.54$, $p < .001$) and when including a cubic relation between ratings and exploration (estimates for uncertainty: $\beta = 0.082$, $z = 24.07$, $p < .001$). These results provide further evidence that uncertainty per se led customers to explore a restaurant.

Learning and dropout

One concern when analyzing customers' order ratings over time is that an increase in order ratings, i.e., a learning effect, could be due to dropout: if customers who did not like their order dropped out of the service, whereas customers who remained tended to produce higher ratings, then this could also lead to higher ratings over time. However, several observations speak against the learning effect being driven by dropout alone. First, the difference in ratings between people who dropped out of the service at any point in time and people who remained was only small, with an effect size of $d = 0.02$. Secondly, even customers with more than 100 past orders continued to improve their ratings over time, with a mean correlation between the number of past orders and ratings of $r = 0.044$ (99.9% CI: 0.038, 0.050). Finally, we estimated learning curves for all participants using mixed-effects regression, with a random effect of the number of past orders per customer. This regression analysis revealed a significantly positive within-customers effect of number of past orders onto ratings with $\beta = 0.08$ (99.9% CI: 0.006, 0.010).

Learning across and within restaurants

Another question is whether customers get better at exploring new restaurants or at ordering from a particular restaurant over time. We therefore assessed how customers' ratings changed over different explored restaurants over time and how their ratings changed over time when ordering from and returning to their favorite restaurant (most ordered from for each customer individually). Fig. S3 shows that both processes seem to happen at the same time.

Customers explored new restaurants they liked more over time ($r = 0.054$, 99.9% CI: 0.047, 0.061; Fig. S3a). Customers also produced higher ratings for their favorite restaurants over time ($r = 0.055$, 99.9% CI: 0.049, 0.061; Fig. S3b). Thus, learning seems to happen both across new and within known restaurants.

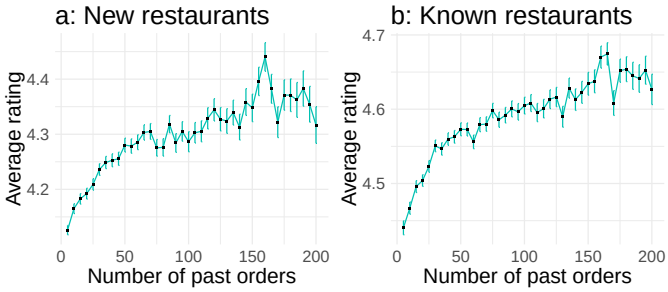

Fig. S3. Learning over newly explored and known restaurants. a: Average order rating for newly explored restaurants by number of past orders. **b:** Average order rating for customers' favorite restaurant by number of past orders.

Learning after initial exploration

We also assessed how participants' ratings improved when ordering from the same restaurant after the initial dip in ratings when exploring a restaurant for the first time. We therefore calculated the mean ratings over time over all customer-restaurant pairs after the first time a customer had explored a restaurant. As shown in Figure S4, exploring a restaurant for the first time led to lower ratings than the average over all orders (4.257, 99.9% CI: 4.250, 4.265; overall mean: 4.324). However, customers improved quickly in their ratings when reordering from the same restaurant ($r = 0.091$, 99.9% CI: 0.085, 0.097) leading to higher than average ratings for all order times greater than 2 (4.521, 99.9% CI: 4.551, 4.532).

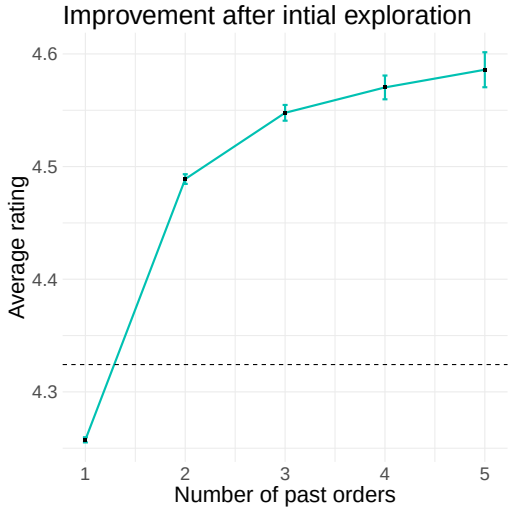

Fig. S4. Learning after initial exploration. The rating for re-visited restaurants increases after the first of exploration. Dashed line indicates the overall mean of restaurant ratings. Standard errors represent the 95% confidence interval of the mean.

RPE, ratings, and reordering from a restaurant

Since RPE is essentially a function of the displayed rating and the provided rating per order, one might also ask which of those is most predictive of customers reordering from a particular restaurant. To assess this, we regressed each of those variables individually onto customers' reordering variable in a mixed-effects logistic regression. This showed the RPE (BIC=483531) was a better predictor for reordering from a restaurant than either the restaurant's mean rating (BIC=485333) or the order rating (BIC=483705).

Chosen feature values over time

We analyze what feature values customers chose on average when exploring new restaurants over time.

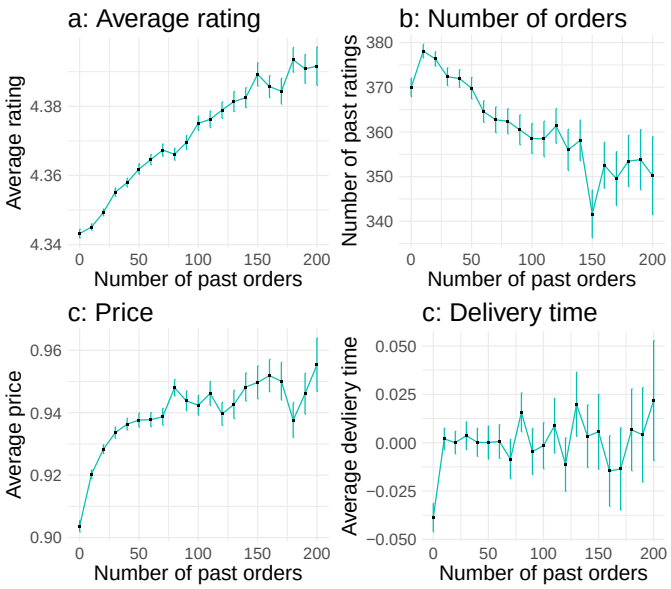

Fig. S5. Average feature values of explored restaurants over time. a: Average displayed rating of explored restaurant. **b:** Average number of past orders of explored restaurant. **c:** Average relative price of explored restaurant. **d:** Average standardized delivery time of explored restaurant.

With experience, customers explored restaurants with higher average ratings ($r = 0.093$, 99.9% CI: 0.086, 0.100; Fig. S5a), fewer past ratings ($r = -0.034$, 99.9% CI: -0.041, -0.027; Fig. S5b), and higher relative price ($r = 0.033$, 99.9% CI: 0.026, 0.033; Fig. S5c). Although the correlation between number of past orders and average delivery time of explored restaurants was practically 0 ($r = 0.007$, 99.9% CI: 0.000, 0.014; Fig. S5d), customers did prefer somewhat faster delivery times for their first 5 orders (mean time= -0.04, 99.9% CI: -0.06, -0.01) than for later orders (mean time= 0.00, 99.9% CI: -0.005, 0.006).

Exploratory change points

To check whether customers' exploration behavior was indeed influenced by their experience or rather mostly by stable individual differences, we tested how they were affected by extreme outcomes. We looked at customers who had experienced an RPE that was either positive and more than two standard deviations better or negative and more than two standard deviations worse than their regular RPEs. We then assessed how likely customers were to explore the cuisine type that had generated the extreme outcome (i.e. the target cuisine type) before and after the outcome had occurred. This analysis showed clear evidence of exploratory change points (see Fig. S6). Although both groups were equally likely to explore the target cuisine type (both $p = 0.187$) before the extreme outcome had occurred, the group experiencing a good outcome was far more likely to explore the target cuisine type after the extreme outcome ($p = 0.232$, 99.9% CI: 0.225, 0.238) than the group experiencing a bad outcome ($p = 0.181$, 99.9% CI: 0.173, 0.181).

Furthermore, the difference in exploration behavior remained significant for more than five of the next exploratory

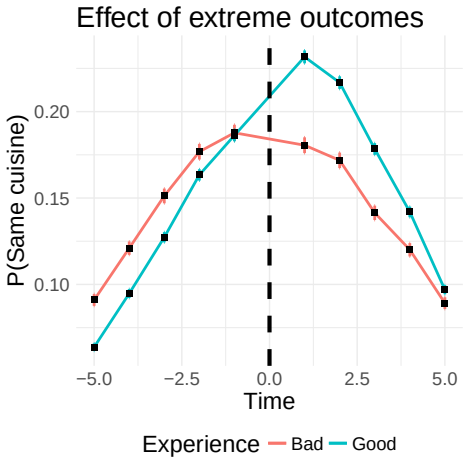

Fig. S6. Change of exploration behavior after very good and very bad outcomes. Probability of exploring the same cuisine type as was explored when an extreme outcome occurred. Extreme outcome occurred a time point $t = 0$, indicated by a dashed line. Standard errors represent the 95% confidence interval of the mean.

purchases (and for more than the next 15 purchases in total), indicating a strong change in cuisine type exploration after extreme outcomes. We therefore conclude that customers can be strongly affected by extreme outcomes leading to exploratory change points that cannot be easily explained by stable individual differences.

Sampling entropy and directed exploration

We calculated customers' mean entropies over the next 4 samples after either a positive or a negative reward prediction error. Shannon's entropy is defined as

$$H = -\sum_i p_i \log p_i, \qquad [15]$$

where i indicates a restaurant within customer's 4 next choices. One of our predictions was that entropy would be higher after negative RPEs than after positive RPEs. We derived this prediction from the fact that a UCB sampling strategy updates both its mean and uncertainty after observing an outcome. 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. We confirmed this prediction in our data. Here, we check if this prediction holds in simulated data that was produced by either the GP-UCB or the GP-M model.

Synthetic data. For a first check of our prediction, we generated synthetic data using samples from a univariate Gaussian Process. Specifically, we created a one-dimensional meshed grid of options with $x \in [0, 0.2, 0.4, \cdots, 10]$. We then sampled a target function from a Gaussian Process with $\lambda = 1$ and optimized this function by using either a softmaximized mean-greedy (GP-M) or a upper confidence bound exploration strategy (GP-UCB) over 20 trials. Moreover, we tracked the models reward prediction error, defined as the difference between its predicted mean for a sampled option and the actual outcome of that option. We repeated this simulation 100 times for both models and afterwards calculated the sampling entropy for the models' next 4 choices after having observed an outcome. Feeding the reward prediction error into a mixed effects

regression with the sampling entropy as the dependent variable and a random intercept for simulation number, we found a significant effect of RPEs onto entropy for the GP-UCB model ($\beta = -0.43$, $SE = 0.01$, $t(1454.5) = -32.70$, $p < .001$), but not for the GP-M model ($\beta = -0.001$, $SE = 0.006$, $t(1454.5) = -0.122$, $p = .9$). Thus, UCB sampling leads to higher entropy after negative outcomes than after positive outcomes, whereas this difference is not pronounced in data generated by a soft-maximizing sampling strategy.

Customer data. In a second analysis, we looked at the sampling entropy difference between the GP-M and GP-UCB models in simulated customer choice data. We focused on the data for the new customers generated for our model comparison. For each customer, we generated a choice set by extracting all sampled restaurants and their features (price, rating, number of ratings, and delivery time). Furthermore, we estimated a utility distribution (the distribution of ratings for each restaurant by customer), by using a hierarchical model with a normal distribution, $\mathcal{N}(4.5, 1)$, as the prior of the mean and a Cauchy distribution, Cauchy$(0, 1)$, as the prior over the variance of the restaurant's utility. The resulting data set can be seen as 3,772 consecutive bandit tasks, where each task contains as many options as unique restaurants a customer had sampled and—for each restaurant—a reward distribution estimated by a hierarchical model based on that customer's ratings. We then let both a GP-UCB and a GP-M model perform within this task, letting them sample as many restaurants as each of the customers had sampled. Afterwards, we calculated the entropy of the next 4 sampling steps as a function of the RPE (Fig. S7).

To assess the effect of RPE on sampling entropy, we regressed—for both sampling strategies individually—the RPE onto the entropy of the next 4 sampling steps in a mixed effects regression while also adding a random intercept for simulated customers. This showed that RPE had a significant effect onto sampling entropy for the UCB sampling strategy ($\beta = 0.031$, $p = .007$) but not for the softmax mean-greedy sampling strategy ($\beta = 0.015$, $p = .07$). We therefore conclude that our theoretical prediction holds in both synthetic and data-driven simulations of exploratory behavior over time.

Clustering analysis

As described in the Materials and Methods, we clustered the 20 most frequent cuisine types appearing within our data set. One appropriate clustering solution of this analysis contained 7 main clusters. The scree plot of the clustering analysis (Fig. S8) confirmed our 7 clusters solution.

Effect of environment analysis

To estimate whether or not customers explored more frequently in cities with higher mean ratings, we calculated the average rating over all restaurants as well as the average exploration rate for every city. The correlation between these two variables was $r = 0.32$, $t(98) = 3.19$, $p = .002$. Even when simultaneously correcting for a city's average restaurant price, order volume, average number of ratings per restaurant, and average number of ratings per customer, the resulting partial correlation was still significant ($r = 0.25$, $t(98) = 2.54$, $p = .02$).

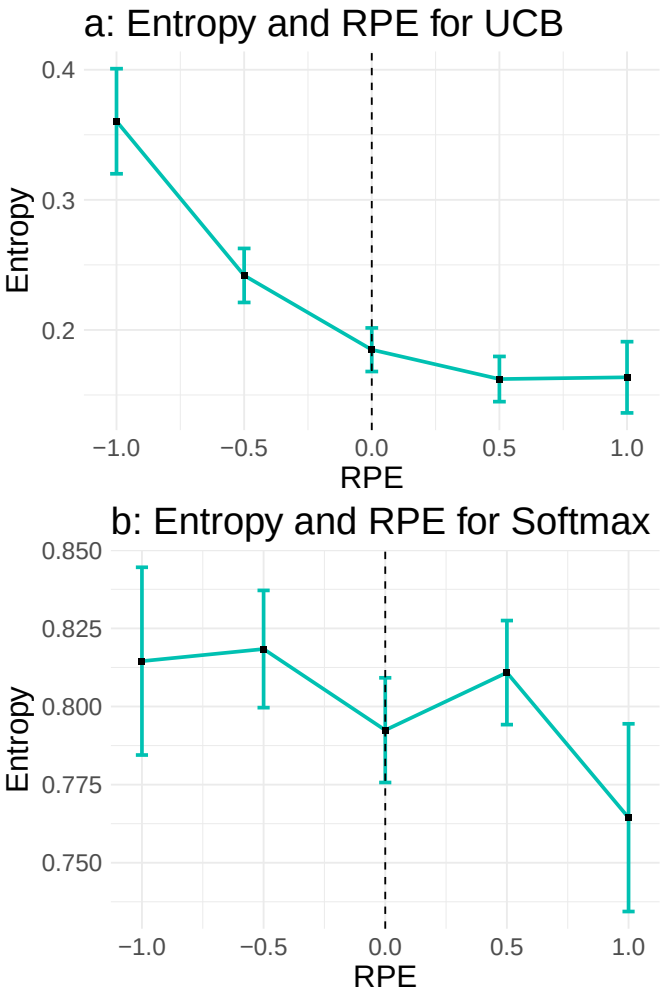

Fig. S7. Entropy and prediction error for UCB and (softmaximized) mean-greedy sampling in a generated restaurant data set.

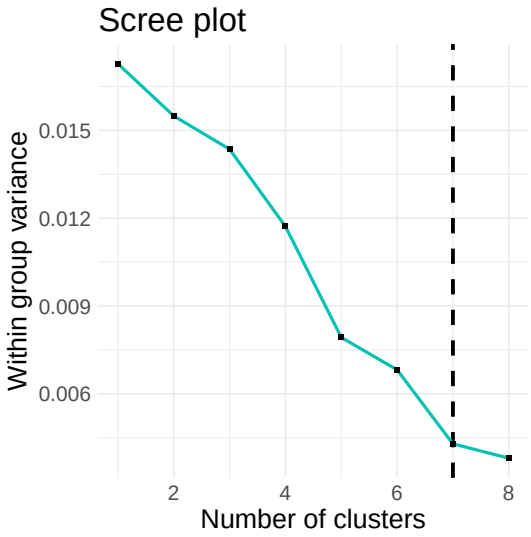

Fig. S8. Scree plot of clustering analysis showing the amount of within variance by number of clusters. Dashed line indicate 7 clusters solution.

Predictability analysis

For the predictability analysis, we assessed—for every city—how predictable customers' ratings were based on the 4 features

used throughout all of our analyses. Doing so, we only used the data set consisting of orders that customers had rated afterwards. We then sampled a learning and a test set for each city consisting of 100 orders each, fitted a linear regression model to the learning set, and used this model to predict customers' order ratings in the test set. Repeating this analysis 100 times for every city revealed how predictable the quality of restaurants within one city was, measured by how well the regression model performed in the test set. We then correlated this predictability measure with the quality of exploratory choices (the mean rating of explored restaurants within a city). This correlation was significantly positive ($r = 0.73$, $t(114) = 10.8$, $p < .001$). Again simultaneously correcting for a city's average restaurant price, order volume, average number of ratings per restaurant, and average number of ratings per customer, the partial correlation remained significant ($r = 0.48$, $t(114) = 5.7$, $p < .001$).

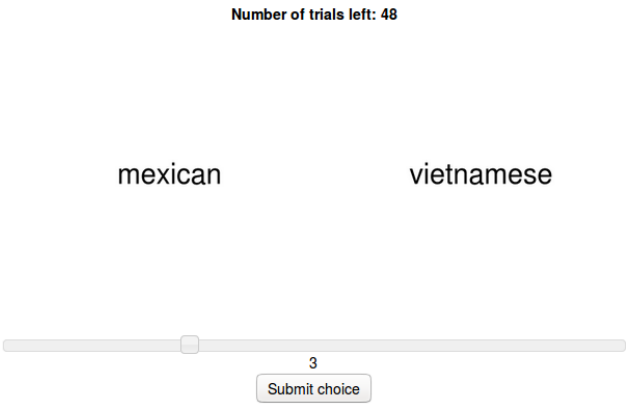

Fig. S9. Screen shot of similarity rating study.

Estimating causal directions

Three of our main theoretical predictions were: (1) customers explore more in cities with higher average ratings; (2) customers explore more successfully in cities where ratings were more predictable; and (3) relative uncertainty has a positive effect on customers' tendency to explore a restaurant. Even though these predictions were derived from past empirical studies, it is nonetheless hard to make strong claims about causal directions in a natural and complex data set. Here, we additionally use another statistical method to estimate causal directions based on observational data proposed by Peters et al. (28). Specifically, this method uses additive noise modeling to assess the residuals when performing a nonlinear regression from one variable to another and vice versa, and then applies kernel independence tests to decide about the causal direction, judging the direction as more likely in which the resulting residuals are more independent. We refer the interested reader to the original paper, but note here that this method gets up to 85% of classifications correct in a very challenging "causal directions benchmark" correct.

We applied this model to the city-specific variables of the mean exploration rate and the average restaurant rating. When regressing the exploration rate onto a city's average restaurant rating using general additive models, this method assessed the probability of independence of the residuals as $p = 0.46$, whereas that probability was $p = 0.58$ the other way around. This method therefore weakly classified the average rating to be more likely the cause of the mean exploration rate than vice versa. We then used this approach to assess the directions of the connection between a city's predictability and customer's exploratory success. Whereas the probability of independence of the residuals was $p = 0.08$ when regressing exploratory success onto predictability, that probability was $p = 0.85$ the other way around. There was thus strong evidence that predictability caused exploratory success according to the causal direction estimation method.

Finally, we analyzed the effect between a restaurant's relative uncertainty and the tendency to be explored. To do this, we estimated every customer's mean relative restaurant uncertainty and mean proportion of exploratory choices. This analysis assessed if customers who explored more did so because of high relative uncertainty in their environment or if customers exploring more often caused higher uncertainties. The resulting p-values for independence were both relatively small as this was a very large data set for both regressions (10^{-10} and 10^{-7}). However, the p-value for independence when regressing exploration onto relative uncertainty was 10^3-times lower than vice versa, showing strong evidence that relative uncertainty led to increased exploration, according to the causal estimation model. Taken together, these results yielded additional evidence that the postulated directions of our predicted and confirmed effects are correct.