Quantifying Curiosity: A Formal Approach to Dissociating Causes of Curiosity

Emily G. Liquin (eliquin@princeton.edu)
Frederick Callaway (flc2@princeton.edu)
Tania Lombrozo (lombrozo@princeton.edu)

Department of Psychology, Princeton University
Princeton, NJ 08540 USA

Abstract

Curiosity motivates exploration and is beneficial for learning, but curiosity is not always experienced when facing the unknown. In the present research, we address this selectivity: what causes curiosity to be experienced under some circumstances but not others? Using a Bayesian reinforcement learning model, we disentangle four possible influences on curiosity that have typically been confounded in previous research: surprise, local uncertainty/expected information gain, global uncertainty, and global expected information gain. In two experiments, we find that backward-looking influences (concerning beliefs based on prior experience) and forward-looking influences (concerning expectations about future learning) independently predict reported curiosity, and that forward-looking influences explain the most variance. These findings begin to disentangle the complex environmental features that drive curiosity.

Keywords: curiosity; learning; surprise; uncertainty; expected information gain

Suppose you are a baker trying to figure out which of two new yeast strains (A or B) will more consistently produce delicious bread. Every night, you bake a loaf of bread with one of the two strains. But there's a problem: one of your employees is careless, and sometimes they fail to clean the kitchen equipment properly. Undetected salt in the bowls contaminates the yeast, and the bread fails to rise properly. As a result, you don't know whether a given loaf's failure is due to the viability of the yeast strain, or to the carelessness of your employee.

After a modest string of successes with strain A, you observe a failure with strain A—how curious are you to know whether your employee failed to clean the bowls properly? After a modest string of failures with strain B, you observe another failure with strain B—how curious are you now? Intuition suggests you may be more curious about your employee's possible negligence in the first case. Why?

Characterizing when and why we experience curiosity is valuable not only because of the link between curiosity and learning (e.g., Kang et al., 2009), but also because it sheds light more broadly on how our mental states, in interaction with our environment, motivate action. In the current research, we introduce a new paradigm for quantifying and dissociating the contribution of several factors to the experience of curiosity. Below, we review theoretical and empirical support for several candidate features that may explain the selectivity of curiosity before introducing our paradigm and two experiments that employ it.

Potential Influences on Curiosity

If the goal of curiosity is to guide an agent towards inquiry that will lead to learning (Berlyne, 1954; Loewenstein, 1994), curiosity should be triggered by expectations about whether learning will occur as a result of pursuing inquiry. For example, if you were to investigate whether your employee cleaned the bowls on a given day, how much would you expect to learn about your employee's actions ("local" anticipated learning, about the target of your question)? And how much would you expect to learn about the viability of the yeast strain ("global" anticipated learning, about your ultimate learning goal)?

Research in neuroscience and developmental robotics has used measures of expected learning to generate theories of curiosity-driven learning (Friston et al., 2017; Oudeyer, Kaplan, & Hafner, 2007) and to test these theories in artificial agents (Macedo & Cardoso, 2012; Oudeyer et al., 2007). One important example is *expected information gain* (Oaksford & Chater, 1994), which quantifies the expected reduction in uncertainty over a hypothesis space after receiving the answer to a particular query. Expected information gain has been used to model goal-directed exploration and question asking (Coenen, Rehder, & Gureckis, 2015; Markant & Gureckis, 2012; Rothe, Lake, & Gureckis, 2018; Ruggeri et al., 2016; Ruggeri, Sim, & Xu, 2017), and recent work suggests that people experience more curiosity about the answer to a question when they expect it to teach them something new and valuable (Liquin & Lombrozo, 2020). Moreover, related work shows that learners experience curiosity about a given stimulus proportional to the probability of encountering that stimulus again in the future (Dubey & Griffiths, 2020).

If an agent is solely motivated to learn an accurate representation of the world,[1] these *forward-looking* evaluations of future learning are likely to be the optimal

[1] Of course, human agents surely have goals beyond just "getting the world right." In fact, prior work has described an optimal policy for curiosity based on the premise that curiosity's goal is to increase an agent's ability to take the right actions (Dubey & Griffiths, 2020). As our experiments do not require participants to take actions, we only consider the goal of learning in the present research.

triggers of curiosity. For example, if your goal is purely to learn about your employee's negligence, local anticipated learning describes the optimal policy for curiosity (assuming curiosity is followed by inquiry behavior). Likewise, if your goal is purely to learn the viability of your yeast strains, global anticipated learning describes the optimal policy for curiosity. However, instead of measuring these forward-looking considerations directly, the bulk of previous empirical research has instead considered the association between curiosity and several *backward-looking* considerations, which compare one's current situation to what one already knows based on prior experience.

For, example, one plausible cause of curiosity is surprise or a violation of expectation (Berlyne, 1966; Stahl & Feigenson, 2015). Being curious after a prediction fails (e.g., that the bread with yeast strain A will rise well) makes sense insofar as the surprise signals some inaccuracy in one's current beliefs, and thus pursuing inquiry will likely lead to future learning. Prior work finds that curiosity can be triggered by the recognition of errors in one's own beliefs (Vogl et al., 2020), and relates the experience of curiosity to prediction errors encoded in the hippocampus and anterior cingulate cortex (Gruber & Ranganath, 2019).

Another plausible cause of curiosity is uncertainty (which may be focused on local or global information, as for anticipated learning). Prior research has shown that local uncertainty is related to curiosity: for example, curiosity about the answer to a trivia question is related to how confidently the participant believes they know the answer (Dubey & Griffiths, 2020; Kang et al., 2009). Local uncertainty can also be objectively quantified and manipulated to influence curiosity (Kobayashi et al., 2019). However, prior research addressing global uncertainty as it relates to curiosity is sparse (cf. Markant, Settles, & Gureckis, 2016).

These backward-looking considerations (surprise and uncertainty) are likely related to curiosity precisely because they are useful heuristic cues to when learning should occur as a result of inquiry. In the bulk of empirical research on curiosity, and perhaps in many real-world contexts, these potential precursors to curiosity are conflated. For example, in the case where a failure with yeast strain A is observed after a string of successes with the same strain, the failure violates your expectations, leads to local uncertainty about the source of the failure, leads to global uncertainty about the viability of the yeast strain, and leads to anticipated learning about both sources of uncertainty.

Nevertheless, recent research (Liquin & Lombrozo, 2020) suggests that features like surprise, uncertainty, and expected learning explain unique variance in participant ratings of curiosity. Furthermore, more variance in curiosity ratings was explained by expected learning than by surprise or uncertainty, controlling for other possible precursors to curiosity. In these studies, however, surprise, uncertainty, and expected learning were based on participant report, so it is not clear what mathematical quantities these self-reports might track. In another study (Markant & Gureckis, 2014), forward- and backward-looking considerations were quantitatively disentangled, but these quantities were used to describe exploration in the service of an experimenter-provided learning goal, rather than intrinsically motivated curiosity (see Gottlieb & Oudeyer, 2018). Additionally, this research only considered local uncertainty and global anticipated learning, and it did not consider whether multiple triggers could operate simultaneously. Thus, in the present research, we ask: Does curiosity track optimal forward-looking features and/or heuristic backward-looking features? And if both sets of triggers are at play, which explain the most variance in curiosity?

The Present Research

We present a paradigm and computational model that allows us to quantify and dissociate several factors that may drive curiosity in a value-learning context. Again, this is challenging because these quantities are often highly associated. In our example of baking—and in paradigms that have been used in prior work (Baranes, Oudeyer, & Gottlieb, 2015; Gruber, Gelman, & Ranganath, 2014; Kang et al., 2009; Kobayashi et al., 2019; Liquin & Lombrozo, 2020)—these quantities hang together and cannot be easily disentangled.

In our paradigm (based on Dorfman et al., 2019), participants observe a series of choices between two mines, which produce either rocks or gold. Outcomes (rocks or gold) depend in part on the underlying reward distribution of the mines (i.e., how likely the mine is to produce gold). However, outcomes are occasionally influenced by a latent variable: a bandit who steals the gold from both mines and leaves only rocks. Thus, a negative outcome (rocks) can be produced by either the mine's failure to supply gold or by the bandit's intervention. Participants rate their curiosity about whether the bandit intervened after observing each outcome. Using a Bayesian reinforcement learning model, we quantify four possible drivers of curiosity in the task: surprise, local uncertainty (which is also local expected information gain), global uncertainty, and global expected information gain. To preview our results, we find that all four features correlate with curiosity, but that the two forward-looking anticipated learning features explain the most unique variance.

Computational Model

To test the hypothesized determinants of curiosity, we extract multiple quantities from a Bayesian reinforcement learning model that tracks the probability of receiving rocks or gold from each of the two mines.[2] For ease of exposition, we present the model for a single mine, noting that it is extended to two mines by separately tracking the evidence about each mine with a copy of the single-mine model.

[2] The model is similar to that presented in Dorfman et al. (2019), but it represents exact posterior distributions over each mine's reward probability, rather than an approximate posterior mean. This is necessary to derive the information-theoretic features.

The generative model is as follows. On each trial, t, the selected mine produces gold with probability θ. However, with probability ϵ, the bandit intervenes, stealing the gold. This intervention is modeled as a latent variable, $Z_t \sim \text{Bernoulli}(\epsilon)$, where $Z_t = 1$ means the bandit intervened on trial t. Gold is only received when the mine produces it and the bandit does not steal it, thus the reward for the trial is distributed $R_t \mid \theta, Z_t \sim \text{Bernoulli}(\theta(1 - Z_t))$.

We assume that ϵ is known (participants are told its value), but that θ must be estimated. We further assume a uniform prior on θ between 0 and 1. Unfortunately, the standard Beta-Bernoulli process does not apply in this case because of the bandit interventions, Z_t. However, the posterior can still be numerically computed by marginalizing over the sequence of interventions, $\vec{z}$,

$$p(\theta \mid r_{1:t}) \propto \sum_{\vec{z}} \prod_{t} \text{Bernoulli}\,(z_t; \epsilon)\text{Bernoulli}(r_t; \theta z_t).$$

Although marginalizing over $\vec{z}$ directly is intractable, this quantity can be efficiently computed by noting that it depends only on the number of times each possible combination of (r, z) occurs. We omit this derivation for reasons of space. We compute the normalizing constant by adaptive quadrature.

On each trial, we extract four quantities:

1. Reward prediction error (RPE) magnitude encodes the extent to which the observed reward on a given trial differs from the expected reward (i.e., surprise). It is defined $|R_t - \overline{r}_t|$, where we consider two definitions of the expected reward, $\overline{r}_t$. First, the true Bayesian expected reward is defined $\overline{r}_t^{\text{bayes}} = E[R_t | r_{1:t-1}] = p(Z = 0)\overline{\theta}_t$, where $\overline{\theta}_t = \int_0^1 \theta\, p(\theta \mid r_{1:t-1})\, d\theta$ is the posterior mean estimate of θ *before* seeing the current reward. We additionally consider a biased estimate that does not take into account the possibility of intervention, $\overline{r}_t^{\text{bias}} = \overline{\theta}_t$. Note that the latter most closely corresponds to the RPE in Dorfman et al.'s (2019) model of the task.

2. Local uncertainty/Local expected information gain (EIG) encodes both uncertainty and anticipated learning about the target of curiosity, which in our case is the presence of the bandit, z_t. We define local uncertainty as the entropy of the predictive distribution, $H(Z_t \mid r_{1:t}) = -\overline{z}_t \log \overline{z}_t - (1 - \overline{z}_t)\log(1 - \overline{z}_t)$ where $\overline{z}_t = p(Z_t = 1 \mid r_{1:t}) = \int_0^1 p\,(Z_t = 1 \mid \theta, r_t) p(\theta \mid r_{1:t})\, d\theta$ is the conditional probability of the bandit having intervened on this trial given the just-observed reward and the history of previous rewards. We define local EIG as the expected reduction in entropy in the predictive distribution, $H(Z_t \mid r_{1:t})$, after observing the value of z_t. Because observing the value of z_t would reduce entropy to zero, local EIG is exactly equal to local uncertainty.

3. Global uncertainty describes uncertainty about variables other than the immediate target of curiosity. Here, we focus on uncertainty about the reward probability of the selected mine. This is defined $H(\theta \mid r_{1:t}) = -\int_0^1 p\,(\theta \mid r_{1:t})\log(p(\theta \mid r_{1:t}))$.

4. Global expected information gain (EIG) describes expected learning about the reward probability of the selected mine. Formally, it is the reduction in entropy of the posterior distribution of θ after observing the value of Z_t, $\sum_{z \in \{0,1\}} p\,(Z_t = z \mid r_{1:t}) H(\theta \mid r_{1:t}, Z_t = z) - H(\theta \mid r_{1:t})$.

While RPE magnitude and global uncertainty are backward-looking in nature (concerning beliefs based on prior experience), global EIG is forward-looking (concerning expectations about future learning). Local uncertainty and local EIG are equivalent, and thus confound forward- and backward-looking considerations. While RPE magnitude, local uncertainty/EIG, global uncertainty, and global EIG are related to each other, as a mathematical consequence of the model, they can be partially disentangled.

Experiment 1

Method

Participants Participants were 101 adults (ages 20-71, *M* = 36) recruited from Amazon Mechanical Turk to complete an ~11-minute study in exchange for $1.40. Participation was restricted to MTurk workers in the United States who had completed at least 1000 prior tasks with a minimum approval rating of 99%. All participants successfully passed two attention checks. Three additional participants who rated their curiosity as zero on all trials were excluded.

Procedure Participants were instructed that they were supervising a team of miners, with the goal of learning how to get as much gold as possible from a set of two mines. On each trial, participants would observe their miners dig in one of two mines, then observe the outcome (rocks or gold). Participants were told that a bandit intervened to steal the gold from both mines on 30% of trials, so they would receive rocks on those trials. Participants were required to correctly answer two comprehension questions. During the main task, after observing each mine choice and outcome, participants dragged a slider to rate their curiosity: "How curious are you about why this outcome occurred (i.e., about whether or not the bandit intervened)?" The slider ranged from 0 (not at all curious) to 100 (very curious) and its initial position was set to 50. After 50 trials, participants answered two multiple-choice attention check questions. Finally, participants provided their gender and age.

Stimuli Participants were randomly assigned to one of four sequences of 50 choices and outcomes. These sequences were generated randomly with the following constraints: (1) one mine was selected on approximately 60% of trials; (2) on exactly 30% of trials, an agent intervened and rocks were received; (3) on the remaining 70% of trials, a reward was drawn based on the underlying probability θ^c of the chosen mine, with θ^c set at 30% for one mine and 70% for the other mine. Across these sequences, the bivariate Pearson correlations between the four quantities of interest was moderate (ranging between -0.61 and 0.70).

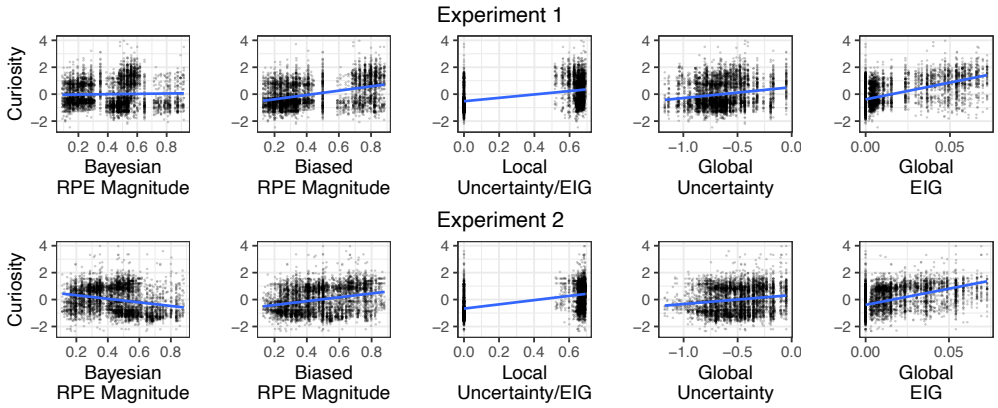

Figure 1: Relation between each quantity and curiosity (z-scored within participants). Each point represents one curiosity rating from a given participant on a given trial. For ease of visualization, curiosity values greater than 4 or less than -2.5 are not plotted (0.2% of all data).

Results

For all analyses, curiosity was z-scored within participants, so that scores reflected the curiosity experienced by a given participant relative to their baseline level of curiosity and its variance across the task.

First, we examined the relationship between curiosity and RPE magnitude, local uncertainty/EIG, global uncertainty, and global EIG. For RPE magnitude, we considered both the true Bayesian RPE that considers the probability of intervention, and the biased RPE that only considers the inferred reward probability of the chosen mine. A series of linear regression models were fit, with each quantity as a predictor in a single model (Figure 1). Curiosity was positively associated with Bayesian RPE magnitude, β = 0.03, 95% CI [0.001, 0.06], $t(5048) = 2.06$, $p = .04$ (R^2 = .001), biased RPE magnitude, $\beta = 0.39$, 95% CI [0.37, 0.42], $t(5048) = 30.29$, $p < .001$ ($R^2 = .15$), local uncertainty/EIG, $\beta = 0.40$, 95% CI [0.38, 0.43], $t(5048) = 31.12$, $p < .001$ (R^2 = .16), global uncertainty, $\beta = 0.17$, 95% CI [0.15, 0.20], $t(5048) = 12.54$, $p < .001$ ($R^2 = .03$), and global EIG, β = 0.55, 95% CI [0.53, 0.58], $t(5048) = 47.08$, $p < .001$ (R^2 = .31). Interestingly, the biased RPE magnitude measure was more strongly related to curiosity than the Bayesian RPE magnitude measure. However, the correlations between biased RPE magnitude and local uncertainty/EIG ($r = 0.11$), global uncertainty ($r = 0.27$), and global EIG ($r = 0.70$) were numerically larger than the same correlations for Bayesian RPE magnitude (r = -0.61, r = 0.21, and r = 0.19, respectively). Thus, the stronger link between biased RPE magnitude and curiosity may be accounted for by its stronger association with other predictors of curiosity.

Subsequently, we tested whether each quantity explained unique variance in curiosity. To do so, we fit two regression models predicting curiosity with all four quantities as predictors. In one regression model, Bayesian RPE magnitude was used to capture surprise, while biased RPE magnitude was used in the second model. In the first model, Bayesian RPE magnitude, β = 0.13, 95% CI [0.08, 0.17], $t(5045) = 5.92$, $p < .001$, local uncertainty/EIG, β = 0.29, 95% CI [0.24, 0.33], $t(5045) = 11.90$, $p < .001$, and global EIG, β = 0.38, 95% CI [0.34, 0.42], $t(5045) = 19.18$, $p <$ .001, all explained unique variance in curiosity. The effect of global uncertainty was marginally significant, β = 0.02, 95% CI [-0.001, 0.05], $t(5045) = 1.90$, $p = .06$. The model explained 33% of the variance in curiosity.

In the second model (Figure 2), biased RPE magnitude, β = 0.10, 95% CI [0.07, 0.14], $t(5045) = 5.85$, $p < .001$, local uncertainty/EIG, β = 0.20, 95% CI [0.17, 0.23], $t(5045)$ = 13.82, $p < .001$, global uncertainty, $\beta = 0.03$, 95% CI [0.004, 0.05], $t(5045) = 2.27$, $p = .02$, and global EIG, β = 0.37, 95% CI [0.33, 0.41], $t(5045) = 18.32$, $p < .001$, all explained unique variance in curiosity. This model also explained 33% of the variance in curiosity.

These results suggest that surprise, local uncertainty/EIG, and global EIG, as captured by an optimal model of learning, all explain unique variance in participant reports of curiosity. Global uncertainty, while positively related to curiosity in isolation, explained less variance in curiosity (if any) when accounting for the other measures. When accounting for all other measures, there was little difference in the extent to which Bayesian RPE magnitude and biased RPE magnitude predicted curiosity. Regardless of how RPE is defined, however, the EIG features explained most of the unique variance in curiosity (32.5% for a model with only EIG features vs. 33% for the full model).

We calculated variance inflation factors (VIFs) for the simultaneous regression models above. The VIF estimates how much the variance of a given coefficient is inflated due to multicollinearity between predictors in a model, and thus can inform whether the null effect found for global uncertainty is a Type II error. For the model where global uncertainty did not reach significance, the VIF for global uncertainty was 1.12 (a VIF over 10 is taken as indicative of serious multicollinearity). This indicates that the null result is unlikely to be due to multicollinearity.

Discussion

The results of Experiment 1 reveal several insights about the

determinants of curiosity. First, surprise, local uncertainty/EIG, global uncertainty, and global EIG can be operationalized and partially separated based on a Bayesian reinforcement learning model. Moreover, unique variance in curiosity is explained by these measures (with the possible exception of global uncertainty). RPE magnitude and global uncertainty are purely backward-looking in nature, while global EIG is purely forward-looking in nature—allowing us to conclude that both forward- and backward-looking considerations are associated with the experience of curiosity.

While these findings are suggestive, the sequences of choices and outcomes presented to participants have some properties that may have made them poorly suited to test the role of these measures in predicting curiosity. In particular, all sequences (unintentionally) began with definitive evidence that one mine had a reasonably high probability of reward and the other had a lower probability of reward. As a result, there were few trials in which global uncertainty was high, and there were few trials in which biased RPE magnitude and global EIG were moderate (Figure 2). We thus attempted to replicate our results in Experiment 2 using a new set of observations and rewards, which capture a fuller and more naturalistic range of these measures of interest.

Experiment 2

In Experiment 2, we presented participants with sequences of observations sampled from the participant data from Experiment 1 of Dorfman et al. (2019). In this prior experiment, participants' task was to select between the two mines on each trial. Each time gold was received from a mine, the participant's bonus payment increased by a fixed amount; each time rocks were received from a mine, the participant's bonus payment decreased by the same fixed amount. Thus, participants' goal was to learn about the mines in order to maximize long-run rewards. In addition to providing a more complete range of model estimates across all trials, this has the additional benefit of providing us with more ecologically valid data sequences (generated by humans rather than sampled randomly).

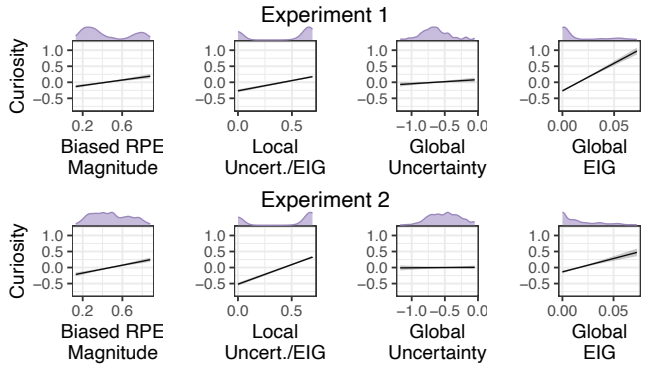

Figure 2: Partial regression plots showing the association between each quantity (using biased RPE magnitude) and curiosity (z-scored within participants), holding the other quantities fixed at their means. The marginal density of the quantity of interest is depicted above each plot.

Method

Participants Participants were 99 adults (ages 22-62, M = 36) recruited from MTurk, as in Experiment 1. Four additional participants were excluded for failing to pass two attention checks. Two additional participants who rated their curiosity as zero on all trials were excluded.

Procedure The procedure matched that of Experiment 1.

Stimuli Participants were randomly assigned to one of eight sequences of choices and outcomes. Using the data from Dorfman et al. (2019), we selected eight sequences of choices and outcomes generated by the 72 participants in the adversarial condition (where a bandit was the intervening agent; there were also two additional intervening agents in their experiments) of their Experiment 1. The four quantities of interest were modestly correlated across these sequences (ranging between -0.70 and 0.64).

Results

As in Experiment 1, we first examined the association between curiosity and each of the four quantities of interest, with each quantity serving as a predictor in a separate regression model (Figure 1). Curiosity was positively associated with biased RPE magnitude, β = 0.27, 95% CI [0.25, 0.30], $t(4948) = 20.06$, $p < .001$ ($R^2 = .08$), local uncertainty/EIG, β = 0.52, 95% CI [0.49, 0.54], $t(4948) = 42.59$, $p < .001$ ($R^2 = .27$), global uncertainty, β = 0.15, 95% CI [0.12, 0.17], $t(4948) = 10.40$, $p < .001$ ($R^2 = .02$), and global EIG, β = 0.47, 95% CI [0.45, 0.50], $t(4948) = 37.87$, $p < .001$ ($R^2 = .22$). However, curiosity was negatively associated with Bayesian RPE magnitude, β = -0.23, 95% CI [-0.26, -0.20], $t(4948) = -16.55$, $p < .001$ ($R^2 = .05$). Again, the contrast in results between the biased and Bayesian measures of RPE magnitude may be accounted for by the correlations between the RPE measures and the other quantities of interest. While biased RPE magnitude was positively associated with local uncertainty/EIG (r = 0.19), global uncertainty (r = 0.23), and global EIG (r = 0.56), Bayesian RPE magnitude was negatively associated with both local uncertainty/EIG (r = -0.70) and global EIG (r = -0.13), two of the strongest predictors of curiosity.

Subsequently, we fit two regression models including all four measures as predictors (one using Bayesian RPE magnitude as a measure of surprise, and the other using biased RPE magnitude). In the model using Bayesian RPE magnitude, all measures but global uncertainty explained significant variance in curiosity; Bayesian RPE magnitude: β = 0.16, 95% CI [0.12, 0.20], $t(4945) = 7.88$, $p < .001$, local uncertainty/EIG: β = 0.53, 95% CI [0.48, 0.58], $t(4945) = 24.48$, $p < .001$, global EIG: β = 0.16, 95% CI [0.12, 0.20], $t(4945) = 8.29$, $p < .001$. This model explained 31% of the variance in curiosity. Notably, when controlling for the other quantities, the association between Bayesian

RPE magnitude and curiosity was positive.

In the model using biased RPE magnitude (Figure 2), a similar pattern of results was found; biased RPE magnitude: β = 0.11, 95% CI [0.08, 0.14], $t(4945) = 7.58$, $p < .001$, local uncertainty/EIG: β = 0.40, 95% CI [0.36, 0.43], $t(4945) = 24.80$, $p < .001$, global EIG: β = 0.16, 95% CI [0.12, 0.20], $t(4945) = 8.11$, $p < .001$. This model explained 31% of the variance in curiosity.

Again, the strength of the association between each of the two measures of RPE magnitude and curiosity was similar when controlling for all other quantities. Additionally, the EIG features explained the majority of the variance in curiosity ratings (30% for a model with only EIG features vs. 31% for the full model). The VIF for global uncertainty was 1.16 in both models, suggesting that its failure to reach significance was unlikely to be due to multicollinearity.

Discussion

Experiment 2 replicates many of the results of Experiment 1, using a naturalistic set of human-generated choice data that captures a fuller range of the model-predicted quantities. Again, we found that RPE magnitude, local uncertainty/EIG, global uncertainty, and global EIG were significantly related to curiosity in isolation. Together, all but global uncertainty—also a weak predictor of curiosity in Experiment 1—explained unique variance in curiosity.

General Discussion

In the present research, we developed a paradigm that quantifies and dissociates the influence of surprise, local uncertainty/EIG, global uncertainty, and global EIG on curiosity. We found that curiosity was independently associated with both RPE magnitude (a backward-looking consideration) and global EIG (a forward-looking consideration) in both Experiments 1 and 2. Additionally, local uncertainty/EIG, which conflates backward- and forward-looking evaluations was a strong predictor of curiosity. Interestingly, and in contrast with previous research on goal-directed exploration (Markant & Gureckis, 2014), forward-looking considerations (local/global EIG) explained a far larger proportion of the variance in curiosity than did backward-looking considerations. However, local uncertainty/EIG conflates forward- and backward-looking considerations, and thus it remains to be determined whether its association with curiosity is accounted for by the forward- or backward-looking component.

Our model does not explicitly define an optimal policy for curiosity, but rather defines an optimal policy for learning. However, under the assumption that a curious agent is motivated to learn, local EIG and global EIG define potential optimal policies for curiosity. Despite the success of these measures in describing participants' curiosity, it remains possible that curiosity is best described by another goal that was not considered in the present research. For example, recent research (Dubey & Griffiths, 2020) has proposed that a rational curious agent should be curious about the stimuli in their environment that maximally increase the value of their total body of knowledge for facilitating future action. Though participants in our task were not responsible for any action, they may have nonetheless taken their goal to be future choices between the two mines, in which case local/global EIG are not necessarily the optimal policies for curiosity. Whether a corresponding policy describes curiosity ratings better than local or global EIG is a question for future research.

Several additional limitations of these studies must be addressed. First, it was not possible to manipulate RPE magnitude, local uncertainty/EIG, global uncertainty, and global EIG completely independently because, while the correlations between these variables were moderate, the variables are non-linearly related. Thus, we cannot make strong claims about the causal influence of any single variable. It also remains an open question whether and when these four precursors to curiosity come apart in real-world contexts. Additionally, participants did not have the opportunity to make choices (i.e., dig in a particular mine) or to attempt to satisfy their curiosity (e.g., look for signs that the bandit had intervened), both of which would likely inform the experience of curiosity.

Finally, we have not considered all possible precursors of curiosity in the present research. In particular, there are several definitions of surprise (for a review, see Munnich, Foster, & Keane, 2019): surprise may be determined by the presence of a low probability event (Reisenzein, 2000), the extent to which one has difficulty explaining an event (Foster & Keane, 2015; R. Maguire, Maguire, & Keane, 2011) or by the extent to which one sees patterns where they expect noise (P. Maguire et al., 2018). In addition, we have not captured all definitions of curiosity. While we focus on state curiosity in response to external stimuli, other research has explored individual differences in trait curiosity (e.g., Litman, 2008). It remains an open question whether participants in our task vary in trait-level curiosity, and whether this variation affects the determinants of state-level curiosity.

Despite these limitations and open questions, the present research introduces a novel paradigm for studying curiosity and identifies several partially separable influences on curiosity. Curiosity was predicted by both backward-looking influences (concerning beliefs based on prior experience) and forward-looking influences (concerning expectations about future learning). These findings help reduce our uncertainty about curiosity, while pointing to new research directions where we can gain additional information.

Acknowledgements

We thank the Concepts and Cognition Lab at Princeton University for their helpful comments. This work was supported by an NSF Graduate Research Fellowship to EL [DGE 1656466]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Baranes, A. F., Oudeyer, P.-Y., & Gottlieb, J. (2015). Eye movements reveal epistemic curiosity in human observers. *Vision Research*, *117*, 81–90.

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology*, *45*, 180–191.

Berlyne, D. E. (1966). Curiosity and exploration. *Science*, *153*(3731), 25–33.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Dorfman, H. M., Bhui, R., Hughes, B. L., & Gershman, S. J. (2019). Causal inference about good and bad outcomes. *Psychological Science*, *30*(4), 516–525.

Dubey, R., & Griffiths, T. L. (2020). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, *127*(3), 455–476.

Foster, M. I., & Keane, M. T. (2015). Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty. *Cognitive Psychology*, *81*, 74–116.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49.

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758–770.

Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. *Neuron*, *84*(2), 486–496.

Gruber, M. J., & Ranganath, C. (2019). How curiosity enhances hippocampus-dependent memory: The Prediction, Appraisal, Curiosity, and Exploration (PACE) framework. *Trends in Cognitive Sciences*, *23*(12), 1014–1025.

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, *20*(8), 963–973.

Kobayashi, K., Ravaioli, S., Baranès, A., Woodford, M., & Gottlieb, J. (2019). Diverse motives for human curiosity. *Nature Human Behaviour*, *3*, 587–595.

Liquin, E. G., & Lombrozo, T. (2020). A functional approach to explanation-seeking curiosity. *Cognitive Psychology*, *119*, 101276.

Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences*, *44*(7), 1585–1595.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98.

Macedo, L., & Cardoso, A. (2012). The exploration of unknown environments populated with entities by a surprise–curiosity-based agent. *Cognitive Systems Research*, *19*, 62–87.

Maguire, P., Moser, P., Maguire, R., & Keane, M. T. (2018). Seeing patterns in randomness: A computational model of surprise. *Topics in Cognitive Science*, (11), 103–118.

Maguire, R., Maguire, P., & Keane, M. T. (2011). Making sense of surprise: An investigation of the factors influencing surprise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 176–186.

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive Science*, *40*(1), 100–120.

Markant, D., & Gureckis, T. (2012). Does the utility of information influence sampling behavior? *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, (34), 719–724. Austin, TX: Cognitive Science Society.

Markant, D., & Gureckis, T. (2014). A preference for the unpredictable over the informative during self-directed learning. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, (36), 958–963. Austin, TX: Cognitive Science Society.

Munnich, E. L., Foster, M. I., & Keane, M. T. (2019). Editors' introduction and review: An appraisal of surprise: Tracing the threads that stitch it together. *Topics in Cognitive Science*, *11*(1), 37–49.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, *11*(2), 265–286.

Reisenzein, R. (2000). Exploring the strength of association between the components of emotion syndromes: The case of surprise. *Cognition & Emotion*, *14*(1), 1–38.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental Psychology*, *52*(12), 2159–2173.

Ruggeri, A., Sim, Z. L., & Xu, F. (2017). "Why is Toma late to school again?" Preschoolers identify the most informative questions. *Developmental Psychology*, *53*(9), 1620–1632.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Vogl, E., Pekrun, R., Murayama, K., & Loderer, K. (2020). Surprised–curious–confused: Epistemic emotions and knowledge exploration. *Emotion*, *20*(4), 625–641.