Intrinsic Rewards in Human Curiosity-Driven Exploration: An Empirical Study

Alexandr Ten, Jacqueline Gottlieb, Pierre-Yves Oudeyer

Intrinsic Rewards in Human Curiosity-Driven Exploration: An Empirical Study

Alexandr Ten (alexandr.ten@inria.fr)
INRIA, Bordeaux, France

Jacqueline Gottlieb (jg2141@columbia.edu)
Columbia University, New-York, USA

Pierre-Yves Oudeyer (pierre-yves.oudeyer@inria.fr)
INRIA, Bordeaux, France

Abstract

Despite their apparent importance for the acquisition of full-fledged human intelligence, mechanisms of intrinsically motivated autonomous learning are poorly understood. How do humans identify useful sources of knowledge and decide which learning situations to approach in the absence of external rewards? While the recognition of this important problem has grown in psychological sciences over the recent years, an intriguing proposition for the possible mechanism comes from artificial intelligence, where efficient autonomous learning is achieved by programming agents to follow the heuristic of maximizing learning progress (LP) during exploration. In this study, we set out to examine the empirical evidence for this idea. Using computational modeling, we demonstrate that humans show signs of following LP while they freely explore and practice a set of multiple learning activities of varying difficulty, including an activity that is impossible to learn. Different approaches to operationalizing the notion of LP and their plausibility in light of empirical data are also discussed. We also show that models combining several types of intrinsic rewards fit better human exploration data than single component models considered so far in theoretical accounts.

Keywords: intrinsic motivation; curiosity; learning progress; computational modeling; model comparisons;

Introduction

Intrinsically-motivated learning is a major force in human development (Oudeyer & Smith, 2016; Gopnik, 2020). Also known as curiosity-driven learning, it allows humans to efficiently explore a variety of experiences and enables them to become more knowledgeable and competent, and thus more prepared for the challenges that the future might bring. The ability to autonomously control one's engagement in learning activities is important because the complex world offers our equally complex minds infinitely many tasks to try. These tasks vary in how many resources they require from individual learners and can even be inherently unlearnable. In order to efficiently navigate through the space of intellectual challenges, humans rely on a cognitive machinery for self-regulated learning (Gottlieb, Oudeyer, Lopes, & Baranes, 2013). However, the precise mechanistic principles underlying intrinsically-motivated learning in humans – specifically, the mechanisms of task selection – remain insufficiently understood (Oudeyer & Smith, 2016; Gottlieb & Oudeyer, 2018).

Related literature on information-seeking in instrumental contexts (Gershman, 2019) proposes that purposeful exploration is guided by reward uncertainty. Exploring uncertain options is traded-off for exploiting rewarding ones with the purpose of potentially increasing the instrumental value in the long-run. A recent study that departs from instrumental decision-making and focuses on information demand for the sake of knowing (Kobayashi, Ravaioli, Baranès, Woodford, & Gottlieb, 2019) reports that humans attribute value to information based on at least two distinct factors. Participants in this study showed individual variability in their tendencies to request information that reduces outcome uncertainty and maximize anticipatory utility.

However, curiosity-driven investigations in humans often span *longer time periods* than most of the literature on intrinsically motivated information-seeking, including (Kobayashi et al., 2019) considers. Such investigations are often concerned with a related but distinct problem of identifying and characterizing sources of information, rather than seeking individual pieces of information (Gottlieb & Oudeyer, 2018). Learning even the basic competences often requires prolonged periods of practice and in the world of many potential things to learn one must constrain the search for information to tasks that are not just unknown but also learnable.

This problem of time-extended exploration is widely recognized in artificial intelligence, where there is a need to explicitly model decision processes over self-imposed goals (Forestier, Portelas, Mollard, & Oudeyer, 2017; Pathak, Agrawal, Efros, & Darrell, 2017). One prominent idea circulates in developmental robotics (Oudeyer, Kaplan, & Hafner, 2007; Colas, Fournier, Chetouani, Sigaud, & Oudeyer, 2019), a field that studies artificial learning in realistic world settings. Here, the agent is provided with means to metacognitively track its competence progress (also called learning progress, or LP) across different tasks (also called goal spaces), and uses this measure to assign interest to tasks that are neither too easy nor too difficult – tasks on which progress in learning is high. In such contexts, agents that maximize uncertainty or incompetence are prone to getting stuck in attempting to learn impossible tasks (Gottlieb et al., 2013).

Effectiveness of autonomous learning in humans raises a possibility that a self-monitoring architecture of the type studied in artificial agents is implemented in the human brain. Yet the empirical support for this idea is sparse. A few related studies provide indirect evidence by reporting human preferences for items and tasks of intermediate difficulty (e.g., Kidd, Piantadosi, & Aslin, 2012; A. F. Baranes, Oudeyer, & Gottlieb, 2014). One recent infant study (Poli, Serino, Mars, & Hunnius, 2020) showed an association between LP and allocation of attention, but did not investigate this link in a

multi-task, time-extended learning setting. Finally, our recent study presented direct evidence for adult humans' reliance on LP computation in a multitask setting with a distractor activity Ten, Kaushik, Oudeyer, & Gottlieb, 2021. However, we only considered one particular form of LP (i.e., absolute competence progress over the recent past in a fixed-size memory). The current study is an immediate extension of our previous work (Ten et al., 2021) investigating the role of LP in a multiple-task setting where learning is self-regulated and time-extended. We use behavioral data from the so-called "internal goal" group of participants from Ten et al., 2021. These participants were required to complete a minimum number of trials (see Methods/Behavioral task) but were otherwise free in a way they wished to interact with the available learning activities. The novelty of the current study over our previous work is mainly in considering a much wider range of decision-making models that are incidentally more expressive.

Our main objective was to assess whether humans rely on LP for guiding their intrinsically-motivated exploration of multiple tasks. For this, we defined two distinct types of LP and used a computational model of decision-making to estimate the individual degree of reliance on these measures. Additionally, we studied the related ideas of competence (Elliot et al., 2000) and uncertainty (Gershman, 2019; Kobayashi et al., 2019) optimization, and used information-theoretic model comparisons to evaluate the involvement of these factors in individual decision-making. Since we set out to fit a separate model to each decision-maker, our secondary goal was to examine the individual differences in self-organized exploration. Finally, as we model the mechanisms of LP computation explicitly, we were also interested in exploring which parameter values for these mechanisms best explain human behaviour.

Methods

Behavioral task

Participants (N = 159), recruited on the Amazon MTurk platform, were paid $1 to perform an online task in which they were allowed to repeatedly choose between four learning activities (Fig. 1). Each trial started by prompting the participant to choose one of 4 activities depicted as families of fictional creatures. After choosing an activity (Fig. 1a, ①), the participant received a randomly drawn exemplar from the corresponding activity, then guessed the category which that exemplar belonged to (categories were strictly nested within activities; Fig. 1a, ②), and after, received binary feedback regarding their guess (Fig. 1a, ③). No additional compensation was provided based on performance during the task.

The only explicit requirement given in the instructions was to complete at least 250 trials of the task. Importantly, the instructions conveyed no expectations for what needed to be done, in what order, or how. The instructions also remained neutral with respect to feedback the participants would receive, stating only that their attempted classification of the stimuli could be correct or incorrect. Before beginning the free exploration task, participants underwent a familiarization stage, which consisted of 4 blocks of 15 trials with randomly sampled exemplars from each activity, such that each block included samples from one activity. After familiarization, participants gave their subjective judgments of the future learnability of each activity and proceeded to complete 250 free-choice trials.

To control the difficulty of the learning activities we manipulated the complexity of their categorization rules. Classification in each activity (except the random activity) was based on the visual features of the exemplars. In the easiest activity (A1), the classes were based on variation along a single variable feature. The next easiest activity (A2) had exemplars that varied along two features only one of which determined the exemplar class. In the most difficult activity which was still learnable (A3), the classification rule was based on a combination between 2 variable features. Lastly, the 4th activity (A4) was essentially unlearnable as feedback provided to the participants was sampled randomly with equally probability of correct and incorrect on each trial. Crucially, we *did not explicitly provide any information about the classification rules of the stimuli*. Participants had to discover the rule of each family (or lack thereof) through trial-end-error and could only achieve high classification accuracy on learnable activities by repeatedly choosing to interact with them.

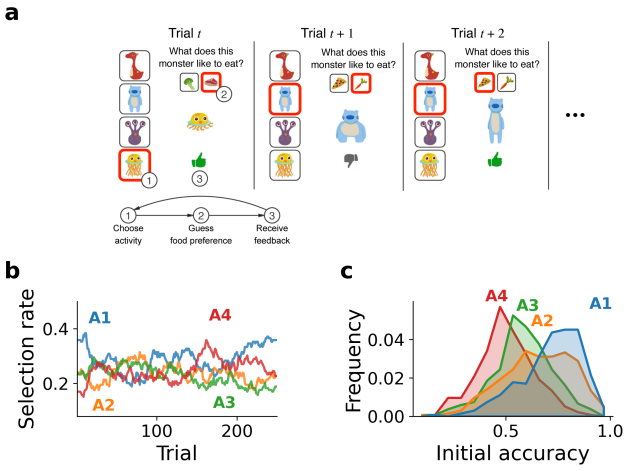

Figure 1: Trial structure during free play. **a**, (adapted from (Ten et al., 2021)). The panels show 3 example trials. Each trial is completed in 3 steps.A trial begins with a choice of the learning activity among the 4 icons of the menu on the left ①. This triggers a presentation of a random exemplar from the chosen stimulus family and a prompt to classify the exemplar into one of two family-specific classes. Upon providing a response ②, the participant receives feedback ③ and proceeds to step ① of the next trial. **b**, Activity selection rates across time were mostly uniform, but a slight preference for A1 can be seen. **c**, Performance, measures as percentage of correct responses over the familiarization trials. The ordering of objective performance is in line with the designed difficulty of the four activities.

Computational modeling

We model activity selection on each trial of self-directed exploration as a value-based probabilistic choice. We assume that the choice of activity partly depends on the subjective utility of the choice options which is itself a function of several factors. We generate a *large* space of hypotheses about these latent factors and assess evidence for these hypotheses using information theoretic criteria. Our hypothesis space was based on a set of four variables corresponding to different types of intrinsic rewards:

- **Competence** (M) defined as the (exponentially weighted) *mean* of correct responses over time; ***Change* in competence** (ΔM) defined as the *difference* between competence over a recent versus distant past (a form of learning progress).
- **Uncertainty** (V) defined as the (exponentially weighted) *variance* of received feedback over time; ***Change* in uncertainty** (ΔV) defined as the *difference* between uncertainty over a recent versus distant past (another a form of learning progress).

Each of these variables represents a separate utility component that is dynamically updated for the sampled activity by receiving feedback, $x_t \in \{0,1\}$, where t is a time index. Competence and uncertainty are computed, respectively, as exponentially-weighted mean (eq. 1) and variance (eq. 2) of the incoming feedback (see Finch, 2009):

$$M_{t|\alpha} = M_{t-1} + \alpha(x_t - M_{t-1}) \tag{1}$$

$$V_{t|\alpha} = (1-\alpha)(V_{t-1} + \alpha(x_t - M_{t-1})^2) \tag{2}$$

Here, $\alpha \in [0,1]$ is a recency-weight parameter that controls the extent to which the latest feedback, x_t, influences the current estimates M and ΔV. We use the "$|\alpha$" notation to indicate that the corresponding quantity is parameterized by α. The change in each of these estimates is computed by taking the absolute difference between the current estimate (respectively, $M_{t|\alpha}$ and $V_{t|\alpha}$) and the estimates $M_{t|c\cdot\alpha}$ and $V_{t|c\cdot\alpha}$ computed with a smaller recency-weight $c \cdot \alpha$, where $c \in [0,1]$. Scaling down the α parameter by c produces estimates that are relatively more representative of the more distant past. The contrast between two estimates representing different time scales provides an estimate of the temporal derivative (or a slope) of the recent estimate. Taking the absolute value of the contrasts[1] causes the utility model to be attracted to positive and negative changes in performance. Therefore, we can compute changes in competence (ΔM) and uncertainty (ΔV) as follows:

$$\Delta M_{t|\alpha,c} = |M_{t|\alpha} - M_{t|c\cdot\alpha}| \tag{3}$$

$$\Delta V_{t|\alpha,c} = |V_{t|\alpha} - V_{t|c\cdot\alpha}| \tag{4}$$

Each of the four definitions provides a formal specification of the space of hypotheses about how a learner might compute a given aspect of his or her performance on a task, and we have to treat them as *hypotheses* because we do not know what the values of the recency-weight parameters might be, i.e., we do not know the time extent of LP computation. However, these are not the only hypotheses we are seeking to evaluate in the study. Our focus is on assessing what combination(s) of the four features – however they are parameterized – best explains how people self-organize their activities. To that end, we define a set of utility functions that define the "interestingness" in each activity as a linear combination of activity features:

$$U_{i,t} = \beta_1 M_{i,t|\alpha} + \beta_2 V_{i,t|\alpha} + \beta_3 \Delta M_{i,t|\alpha,c} + \beta_4 \Delta V_{i,t|\alpha,c} \tag{5}$$

where i indexes a learning activity. Equation 5 represents many different utility models, some of which exclude some or all of the features by settng the corresponding β coefficient to a constant value of 0. Thus, the set of 4 task-performance features gives us 16 hypothesis spaces of variable dimensionality; each hypothesis space is spanned by combinations of free parameters β as well as the recency-weight parameter α and a scaling parameter c where it is applicable.

We model the trial-by-trial choice of a learning activity as a utility-based stochastic process (see Daw, 2011):

$$p_t(\text{choice}_i) = \frac{e^{U_{i,t}\times\tau}}{\sum_{k\in K} e^{U_{k,t}\times\tau}} \tag{6}$$

where $U_{i,t}$ is the subjective value of activity i at time t, and k indexes all items in the set of available activities K (including i); the additional free parameter τ, controls choice stochasticity.

Parameter estimation was performed using the choice and feedback data from individual trials of free play by minimizing the negative sum of log likelihood values over those trials (see Daw, 2011). Parameter optimization was performed using the L-BFGS-B nonlinear numerical optimization method (Byrd, Lu, Nocedal, & Zhu, 1995) with bound constraints of [-1, 1] for the β coefficients, [0, 1] for α and c, and [0, 1000] for τ. We randomly initialized 10 optimization bouts for each participant and each model form and selected the best fitting models for the reported analyses. All code use for modeling and analyses is freely available at `https://github.com/flowersteam/utility-space-cogsci-2021`

Results

Time allocation and learning

As expected, providing minimal cues about the external expectations of the experimenters resulted in a highly variable choice behavior (Fig. 1b). At any point in time participants where approximately equally distributed across the four activities, with occasional peaks of popularity for certain activities (e.g., A1 in the beginning and in the end, A4 around trial

[1] The idea of using the absolute value of the difference comes from the machine learning literature where it was shown to work well for guiding the curiosity-driven exploration (A. Baranes & Oudeyer, 2013; Colas et al., 2019)

160). There was also a lot of variability in the overall percentage of correct responses, but the success rates varied systematically with the designed difficulty of the learning activities (Fig. 1c). A one-way ANOVA of the initial performance measured by % of correct response by activity was significant ($F(3, 1344) = 136.28,\ p < .001$)) and post hoc contrasts using Tukey's HSD correction confirmed that performance was significantly different between any pair of learning activities ($p < .01$). Thus, the difficulty manipulation worked as expected and provided a basis for exploring the set of activities by difficulty.

One potential concern for a behavioral task that empowers the participant to choose their own activities is a potential for unaccounted choice variables. Such potential nuisance effects were minimized by randomizing difficulty levels across monster types: monster families had constant difficulty for each participants, but different families had different difficulties between participants. The benefit of letting the participant decide what to learn when is in mimicking the natural learning environment to better capture the ongoing decision-making process that is concurrent with learning.

Model comparisons

What model forms best explain the choices? We fitted each of the 16 models forms of the variable set to each participant's data (note that the null model that excludes all the utility components corresponds to a uniform-choice model with 0 parameters). The fitted models can be compared within each participant (but not across participants) using the AICc scores (Symonds & Moussalli, 2011). Fig. 2 presents the distributions of AICc scores of all models, grouped by model form (Fig. 2a) and the frequencies with which various model forms had minimum AICc scores within participants. Across participants, the most frequently encountered best model was a trivariate-utility model that included the M ΔM and ΔV components. It was also the model with the lowest median AICc. However, other models forms were also frequently found to be the best models and in general, multivariate models with 3 to 4 components explained participants' choices better than models with less components.

We also compared 4 bivariate model forms with the fixed-size memory competence and learning-progress bivariate model forms from Ten et al., 2021 in order to examine whether fitting the time horizons of competence and LP computations lets us fit the choices better. The fixed-size memory model included 2 utility components: (1) a flat-average competence component computed over the last 15 trials on an activity, and (2) a learning-progress component computed as the absolute difference between flat-average scores of the first one-third of the last 15 trials and the last one-third of these trials (see Ten et al., 2021 for details). The 4 bivariate models from the present study included the following utility compositions: $M + \Delta M$; $M + \Delta V$; $V + \Delta M$; and $V + \Delta V$. One of these model forms had lower AICc than a fixed-size memory model in 65.41% of participants. Compared to any model form from the present study, the bivariate fixed-size memory model provided worse fit in 91.82% of participants.

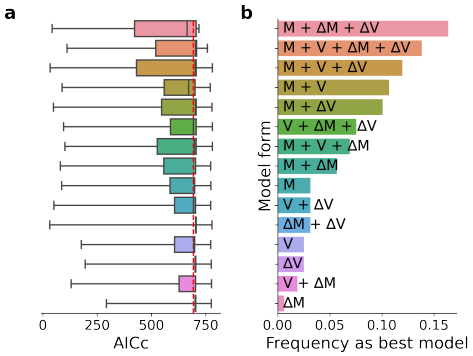

Figure 2: Model comparisons. Different colors identify model forms across the subpanels. **a, Distributions of AICc scores.** The boxes show the interquartile range, with lines inside them representing the median values. The whiskers include the entire range of each distribution. The vertical red line shows the AIC score of the random-choice model (AIC = 693.15). **b, Frequency of the best models**. The plot shows the frequencies of finding each model form as the highest-ranking model within participants.

Whereas criteria like AICc are useful for quantitative assessments of relative likelihoods, it is important to inspect how the models reproduce the relevant patterns in the observed behavior. For this purpose, we simulated freely exploring learners using coefficients from the fitted models to see if the simulated task choices would match the observed choices and in what way might the multivariate models provide a superior fit. We show the results in Fig. 3.

The simulations demonstrate that the best-fitting univariate models could qualitatively reproduce time allocation trends for simpler strategies, like preferring the easiest activity (A1), as seen in the entire sample as well as in the worst 30 learners in our sample[2]. However, the univariate models did not do well at reproducing more sophisticated behaviors, like allocating time in proportion to activity difficulty (as seen in the best 30 learners; Fig. 3, middle row). On the other hand, the best-fitting trivariate models, which included at least one variant of the LP component, were able to reproduce the this pattern well.

Both quantitative and qualitative evaluations of fit of multivariate models strongly suggest that 1) at least one form of LP is present in the most likely models, and 2) activity choices are likely shaped by a combination of different decision factors and no single component can adequately explain self-determined exploration. In the next section, we explore the covariation among the fitted utility-component coefficients across individuals and discuss its relationship with learning.

[2]The best and worst learner groups were determined by the difficulty-weighted %correct score. The weights were assigned by ranking activities by their relative difficulty (i.e., A1 measures were weighted by $\frac{1}{6}$, A2 by $\frac{2}{6}$, and A3 by $\frac{3}{6}$

Examining variability in fitted utility parameters

It is possible that a combination of individually weak components can make up a very strong component. To better understand how people might combine utility components we analyzed the variability of fitted coefficients. Specifically, we used the principal component analysis (PCA) which allowed us to assess how these fitted coefficients varied together in order to identify the main dimensions of variability.

What are the main dimensions of variability across individuals? A preliminary fit showed that 2 components were enough to explain 77% of variation in the fitted coefficients. We used the raw (non-standardized) coefficients, as they were all on the same scale due to the bound-constrains between 0 and 1 on parameter optimization. However, since we used the coefficient values from the best fitting models in each participant, we had to substitute the missing values of the coefficients in models that did not include a certain component with 0, which is consistent with that component not influencing the overall utility in any way.

We present the table of correlations of model coefficients with the fitted PCs (Table 1). The correlations clearly show that variations in ΔV and V coefficients were the most prominent in our participants. The first PC captures the tendency to chose tasks according to the changing uncertainty and to much lesser extents to the estimates of uncertainty and change in competence. The second component is most strongly related to sensitivity to the current estimates of uncertainty, characterizing learners who seek to maximize the feedback variance. This, variation in the two uncertainty-based components seem to explain most of variability among the coefficients of the best-fitting models.

How does the observed variability relate to learning? While most of the best-fitting model forms in Fig. 2 which compares fits by AICc included the M component, the PCA shows that this factor – while important for explaining task choices – did not vary systematically with any of the PCs.

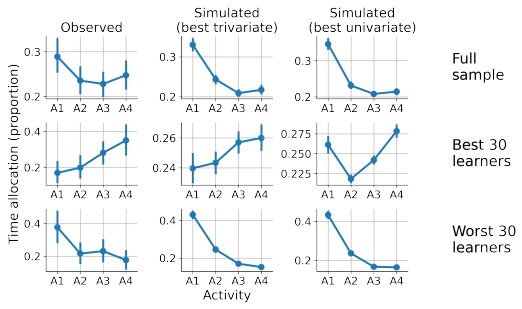

Figure 3: Simulated choices based on fitted parameters. The simulations were performed by first bootstrap-sampling 1000 best-fitting trivariate or univariate models from either the full sample of participants (top row), or 30 best learners (middle row) or 30 worst learners. The bootstrapped models were then used to make free choices for 250 trials using simulated feedback generated with the actual correct-response probabilities observed in the full sample.

	PC1	PC2
Variance explained	.49	.27
ΔV	**.62**	.11
V	.18	**-.46**
ΔM	-.12	-.13
M	.01	.02

Table 1: Correlations of the original variables (fitted coefficients) and two principal components.

Most of variability in the fitted coefficients was attributable to V and ΔV which suggests that participants relied on these signals to varying degrees. We were interested in whether this variation was related to learning.

First, we measured each participant's overall improvement by subtracting (1) the initial difficulty-weighted performance scores computed from the first 15 trials of each learnable task from (2) the final difficulty-weighted scores computed from the last 15 trials on each learnable task. This gave us a measure of the difference between final and initial performance (higher values correspond to more improvement). We then tested whether Pearson correlation coefficients between these scores and each of the PCs was different from 0. The tests showed positive and marginally significant correlation between learning and PC1 ($r(157) = 0.162, p = 0.041$) and a negative correlation between learning and PC2 ($r(157) = -0.158, p = 0.047$), but no correlation between learning any of the individual coefficients except V ($r(157) = 0.211, p = 0.008$). Overall, this shows that *combinations* of components of learning dynamics predict improvement better than individual components. The positive strong correlation between improvement and V is not necessarily at odds with this conclusion, because high feedback variance can indicate changing performance (ΔM) while low feedback variance can indicate either consistently high or low performance (M).

Temporal extent of progress computation

What time-scales for computing LP explain the data best? To answer this question we examined the distributions of the fitted recency-weights taken from the best fitting models of each participant that had either ΔM or ΔV or both in their best model. To simplify the interpretation of the recency-weight parameters, we computed their reciprocals ($n_1 = \frac{1}{\alpha}$ and $n_2 = \frac{1}{c \cdot \alpha}$) which measure, respectively, the extent to which the computations of recent competence and uncertainty depend on older data (n_1), and the extent to which long-term competence and uncertainty estimates depend on it (n_2). Thus, higher values of an n_i parameter can be interpreted as longer time-windows over which a weighted estimate is computed. Estimates computed with lower n_i parameter values are more noisy and less inert.

The top-left histogram of Fig. 4a (note the logarithmic scale) shows the distribution of the fitted n_1 parameters. Most of the values are concentrated at the low end of the distribu-

tion[3], suggesting that the running estimates of competence and uncertainty were computed over relatively small time-windows ($n_1 \leq 10$). The bottom-left plot shows that the contrasting long-term estimates tended to include information about more data from the past as indicated by a less pronounced tail. Nevertheless, the majority of n_2 values were between 1 and 100. The top-right histogram, is also interesting, as it shows the differences in n_1 and n_2 parameters. Corresponding to the distribution of n_2, these spanned up to three orders of magnitude but where mostly contained between 1 and 100. Importantly the same pattern in $n_2 - n_1$ differences was observed in a sub-sample of top 30 learners as measured by their difficulty-weighted final %correct. These distributions show that the fitted models used relatively short time windows for the running estimates of competence and uncertainty, but a range of relatively long and short temporal contrasts to track changes in these estimates. These are in the range in which both the recent estimates and the changes in these estimates are relatively well-behaved (not too noise and not too inert).

[3]Some participants had very small $c \cdot \alpha$ values which resulted in $\log_{10}$ transforms that were too high for further processing. These values where clamped to n_2 derived from the minimum $c \cdot \alpha$ in the data set. This is why there is a stump at the high end of each distribution involving n_2

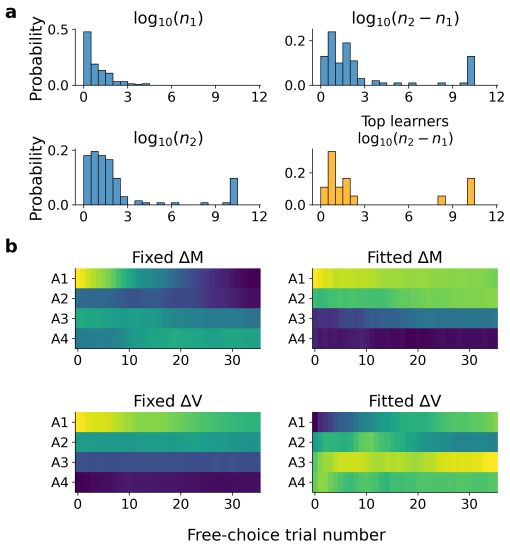

Figure 4: a. Distributions of $\log_{10}$-transformed temporal coefficients, $n_1 = 1/\alpha$ and $n_2 = 1/(c \cdot \alpha)$. The $\log_{10}$ transformation makes the values easy to interpret. Each bin of the histograms has a width of 0.5, so the first bin contains values of n_i approximately between 1 and 3, the next bin has values in [3, 10], then [10, 30], [30, 100], [100, 300] and so on. The bottom-right plot shows the distribution of $\log_{10}(n_2 - n_1)$ scores in the top 30 learners. **b, Comparison of using fixed vs fitted feature computation**. Brighter colors indicate higher relative values (within the panels). In each panel, each pixel shows the value of the labeled quantity for each task, averaged across subjects and for all past trials (i.e., a moving average). Only the best learners where used to generate these plots (top 30 participants in difficulty-weighted final performance across learnable activities).

What forms of LP indicate learnability better? Previous work on modeling free choice in multi-task settings used fixed temporal window sizes for computing running estimates of competence and competence progress (see Ten et al., 2021). We extend this work by explicitly modeling the computation of task features. We are thus in a position to assess the validity of different approaches to operationalizing LP. Ten et al., 2021 used a fixed-size memory approach in which both competence and LP were computed over a moving window of a constant size (15 trials). Specifically, competence was computed as a flat average of correct guesses over a moving window and LP was computed as the absolute difference between the first 10 and the last 9 trials of that window. Moreover, Ten et al., 2021 did not consider any equivalents of V or ΔV but these are easily implemented for fixed-sized windows, so we can compare the validity of LP computation of competence- and uncertainty-based variants as well as the fixed vs. fitted temporal parameterization.

First, we selected the top 30 learners (as measured by their final difficulty-weighted % correct score). It was important to isolate "good learners" because our instructions did not require participants to learn. While some participants underchallenged and seemed to be mainly driven my maximizing percent progress, for the present analysis we were specifically interested in ways to model the role of LP in *motivated* learners. We used this sub-sample's response-feedback histories in each activity to compute the estimates of ΔM and ΔV using both fixed and fitted time-scale parameters. Fig. 4b presents the resulting estimates for the first 50 trials on each activity (starting from trial 16 to accommodate the fixed time windows, which corresponds to the first trial of the free-play stage). The estimates were averaged across participants and smoothed over time to reveal the patterns.

A good measure of LP should reflect changing performance on any task: (1) it should be initially high but decreasing for very easy tasks (A1 and A2), reflecting the initial surge and an eventual plateau, (2) it should also be relatively stable and steady for difficult tasks on which the performance changes slowly (A3), and (3) it should be consistently low for an unlearnable task (A4).

These properties of a good LP measure were observed sporadically across contrasted measures of LP. Fig. 4b shows the initially high but decreasing signal for fixed ΔM, fixed ΔV, and fitted ΔM. A constantly low and unchanging signal was given by fixed ΔV and fitted ΔM. Finally, a relatively high and consistent signal for the intermediate-difficulty task (A3) was provided only by the fitted V measure. Overall, these sporadic patterns might explain why the best models combine multiple features. Moreover, given that the fixed M measure, as used in Ten et al., 2021, does not seem particularly optimal for providing a good LP signal, this analysis also points to the potential benefits of parameterizing and fitting the time-horizon of LP computation, as well as considering different forms of LP.

Thus, the issue of approximating the veridical time-scales over which the LP computation might occur is important if we want to test theoretical predictions about the role of LP in exploration. Therefore, future work addressing this issue more systematically and with more focus will undoubtedly be important for promoting our level of understanding of the mechanisms for self-directed learning.

Conclusion

Our study provides an exploration of the idea that humans rely on LP in organizing free exploration. We present some early evidence for this idea by using a behavioral task that is appropriate for probing information-seeking over extended periods of time. Specifically, we show that humans rely on a combination of decision factors, including how much progress they are making in reducing outcome uncertainty, but also the running estimates of competence and uncertainty. We also show that people rely on a wide range of strategies, but primarily and orthogonally differ in their sensitivities to change in uncertainty and to the level of uncertainty. Finally, we provide some indication that uncertainty progress might be a more efficient form of LP for tracking difficult but learnable tasks while avoiding tasks that cannot be learned.

Future research investigating the ecological roles of different kinds of intrinsic rewards and detailing the principles of their interaction will be useful for advancing our understanding of curiosity-drivel learning in humans.

References

Baranes, A., & Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, *61*(1), 49–73.

Baranes, A. F., Oudeyer, P.-Y., & Gottlieb, J. (2014). The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in neuroscience*, *8*, 317.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, *16*(5), 1190–1208.

Colas, C., Fournier, P., Chetouani, M., Sigaud, O., & Oudeyer, P.-Y. (2019). Curious: intrinsically motivated modular multi-goal reinforcement learning. In *International conference on machine learning* (pp. 1331–1340).

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII*, *23*(1).

Elliot, A. J., Faler, J., McGregor, H. A., Campbell, W. K., Sedikides, C., & Harackiewicz, J. M. (2000). Competence valuation as a strategic intrinsic motivation process. *Personality and Social Psychology Bulletin*, *26*(7), 780–794.

Finch, T. (2009). Incremental calculation of weighted mean and variance. *University of Cambridge*, *4*(11-5), 41–42.

Forestier, S., Portelas, R., Mollard, Y., & Oudeyer, P.-Y. (2017). Intrinsically motivated goal exploration processes with automatic curriculum learning. *arXiv preprint arXiv:1708.02190*.

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, *6*(3), 277.

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, *375*(1803), 20190502.

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

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, *17*(11), 585–593.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, *7*(5), e36399.

Kobayashi, K., Ravaioli, S., Baranès, A., Woodford, M., & Gottlieb, J. (2019). Diverse motives for human curiosity. *Nature human behaviour*, *3*(6), 587–595.

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.

Oudeyer, P.-Y., & Smith, L. B. (2016). How evolution may work through curiosity-driven developmental process. *Topics in Cognitive Science*, *8*(2), 492–502.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning* (pp. 2778–2787).

Poli, F., Serino, G., Mars, R., & Hunnius, S. (2020). Infants tailor their attention to maximize learning. *Science advances*, *6*(39), eabb5053.

Symonds, M. R., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaike's information criterion. *Behavioral Ecology and Sociobiology*, *65*(1), 13–21.

Ten, A., Kaushik, P., Oudeyer, P.-Y., & Gottlieb, J. (2021). Humans monitor learning progress in curiosity-driven exploration. *PsyArxiv*.