An Evaluation of Experimental Sampling Strategies for Autonomous Empirical Research in Cognitive Science

Sebastian Musslick[1,2,3], Joshua T. S. Hewson[1], Benjamin Andrew[1], Younes Strittmatter[1,2], Chad C. Williams[1], George T. Dang[4], Marina Dubova[5], & John G. Holland[4]

[1]Carney Institute for Brain Science, Brown University; [2]Department of Cognitive, Linguistic, and Psychological Sciences, Brown University; [3]Institute of Cognitive Science, Osnabrück University;
[4]Center for Computation and Visualization, Brown University; [5]Cognitive Science Program, Indiana University
Corresponding Author: sebastian@musslick.de

Abstract

In light of constraints inherent to empirical research, such as finite time and resources, there has been growing interest in using artificial intelligence to streamline the scientific process. However, despite advancements in automating scientific discovery, the implementation of strategies for sampling useful experiments remains a challenge. This metascientific study evaluates different experimental sampling strategies based on their effectiveness in advancing the discovery of linear models of human cognition based on synthetic data. We investigate the hypothesis put forth by Dubova et al. (2022) that random sampling of experiments is more effective than model-driven sampling. Indeed, the results of this study indicate that random sampling is more effective in a majority of cases, and that the underperformance of model-driven strategies can be attributed to a narrow sampling of the design space. Despite limitations in our approach, the work presented offers a novel framework for the metascientific study of autonomous empirical research.

Keywords: computational discovery; automated scientific discovery; active learning; metascience

Introduction

Recent work calls for large-scale experiments in cognitive science to facilitate the development of integrative theories of cognition (Almaatouq et al., 2022; Griffiths, 2015). To catalyze such large-scale experimentation, researchers have proposed that artificial intelligence be integrated into the empirical research process (Agrawal et al., 2020; Musslick, 2021; Peterson et al., 2021). However, there is limited knowledge of how to advance the development of statistical models through the effective sampling of experimental conditions. In this metascientific study, we leverage a closed-loop framework for autonomous empirical research to examine the abilities of various experimentation strategies to accurately recover established models of cognition from synthetic data.

Previous efforts to automate sampling of experimental conditions have relied on expert domain knowledge (King et al., 2009; Lindsay et al., 1993) and active learning techniques, which are machine learning methods for selecting the most useful data points given a model (Settles, 2009). In empirical sciences, where an experimental design space has been chosen, active learning can be used to select experimental conditions that maximize the information gained from a novel experiment. In principle, this allows for efficient use of resources, whereby researchers can pinpoint promising conditions rather than exhaustively sample the entire design space.

Recently, active learning approaches were pitted against random sampling and found to be less effective (Dubova et al., 2022). In their metascientific study, Dubova et al. (2022) tasked different theorists, which were represented by autoencoders, to collaboratively reconstruct a ground truth constituted by a mixture of multi-variate Gaussian distributions. Data were sampled using various strategies, including theory falsification, novelty sampling, and random sampling. The results indicated that random sampling outperformed (model-driven) active sampling strategies, presumably because the latter focused on a narrow region of the experimental design space. This finding runs counter to the hypothetico-deductive model and suggests, provocatively, that random sampling may be the optimal way to design experiments and test theories. However, one could argue that the study's setup may not accurately reflect real-world empirical research, which relies, to a large extend, on interpretable, linear statistical models and aims to explain phenomena more complex than multivariate Gaussian distributions.

In this article, we investigate the hypothesis put forth by Dubova et al. (2022)—that random sampling of experimental conditions outperforms active sampling in guiding the recovery of interpretable, statistical models of human cognition from synthetic data. However, unlike in the original study, we simulate the empirical research process with interpretable theories and synthetic datasets, covering paradigms such as object categorization (Luce, 1963; Shepard, 1958), value-based choice (Tversky and Kahneman, 1992), controlled processing (Cohen et al., 1990), and task switching (Yeung and Monsell, 2003). To accomplish this, we leverage a novel framework for autonomous empirical research that relies on computational discovery and closed-loop automation to identify interpretable models of empirical phenomena. We find that, for any given ground truth, random sampling outperforms many active sampling approaches. Our analyses suggest that the benefit of random sampling results from capturing more variance in measurements obtained from the ground truth. However, the most successful sampling strategy depends on the type of data and can be active rather than random. We conclude by discussing the implications for computational discovery in cognitive science and outline future directions for the computational metascience of human cognition.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

Methods

We begin by introducing the test bed used for evaluating different experiment sampling strategies in cognitive science. The test bed relies on an open-source framework for autonomous empirical research, which allows for the integration of mechanisms for model discovery, experiment sampling, and data collection. We then outline analytical procedures for comparing different experiment sampling strategies. The framework is available as the Python package `autora`, and all reported simulations are based on code listed in `https://github.com/autoresearch/autora`.

Overall Approach

Our test bed integrates three software components: (1) an autonomous theorist that constructs quantitative models linking experiment conditions to dependent measures; (2) an autonomous experimentalist that designs novel experiments; and (3) a synthetic environment for data collection. In the current study, we focus on the second component. Specifically, we pair a logistic regression with different experimentalist algorithms and evaluate which combination is best able to recover four models of human cognition that we treat as ground truths.

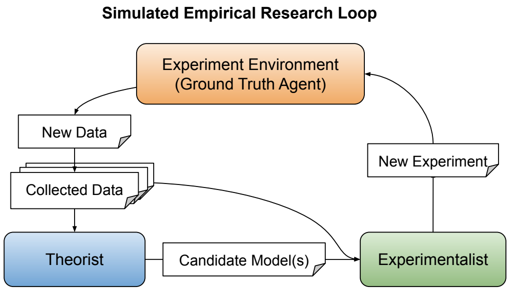

Figure 1: Simulated Empirical Research Loop. An autonomous theorist (blue) generates candidate models based on collected data. The experimentalist (green) proposes novel experiments to be conducted on a synthetic ground-truth agent (orange), from which new data is collected.

We evaluate the effectiveness of each experimentalist in a closed loop of model fitting and experimentation. This loop iteratively executes the following steps: it generates new experimental conditions through the experimentalist, it collects new data through a given ground-truth agent, and fits a logistic regression through the theorist. We validate each fitted regression model by assessing its ability to predict observations obtained from the full space of legal experiment conditions, as defined by the ground-truth agent.

Theorist

We implement an autonomous theorist that uses logistic regression to find a model that correlates experimental variables with noisy dependent measures obtained from a ground-truth agent. We choose a basic logistic regression in this study because (1) it is highly interpretable with respect to the influence of experimental factors on the dependent variable, (2) it is frequently used for modeling choice data, and (3) it can qualitatively recover signature effects of all ground-truth agents[1] For each ground-truth agent, all experimental factors and their interactions are included as regressors, while the observations obtained from each ground-truth agent are considered the regressands. One of the sampling strategies (model comparison) requires two models from the theorist. Thus, we also fitted a second (reference) model that included no interaction terms as regressors.

Experimentalists

We consider the five experimentalists listed in Table 1. The goal of each experimentalist is to identify novel experiment conditions $\vec{x} \in X$, where x_i corresponds to the level of the experiment factor i.

Table 1: Experimentalists. To determine novel experiment conditions $\vec{x}$, experimentalists may use any of the latest candidate models M from the theorist, experimental conditions that have already been probed $\vec{x}' \in X'$, or respective dependent measures $\vec{y}' \in Y'$.

Experimentalist	Function	Arguments		
		M	X'	Y'
Random	$\vec{x}_i \sim U[a_i, b_i]$			
Novelty	$\arg\max_{\vec{x}} \min(d(\vec{x}, \vec{x}'))$		✓	
Least Confident	$\arg\max_{\vec{x}} 1 - P_M(\hat{y}^*\|\vec{x})$, $\hat{y}^* = \arg\max_{\hat{y}} P_M(\hat{y}_i\|\vec{x})$	✓		
Model Comparison	$\arg\max_{\vec{x}} (P_{M_1}(\hat{y}\|\vec{x}) - P_{M_2}(\hat{y}\|\vec{x}))^2$	✓		
Falsification	$\arg\max_{\vec{x}} \hat{\mathcal{L}}(M, X', Y', \vec{x})$	✓	✓	✓

Experimentalists may use information about any of the theorist's candidate models M, experiment conditions that have already been probed $\vec{x}' \in X'$, or respective dependent measures $\vec{y}' \in Y'$. One strategy, referred to as the *random* experimentalist, disregards prior information and instead samples each experiment factor uniformly from a predefined interval, $U[a_i, b_i]$, where a_i and b_i represent the lower and upper bounds of the experiment factor, respectively. Another approach, referred to as the *novelty* experimentalist, selects new experiment conditions that maximize the smallest Euclidean distance, $d(\vec{x}, \vec{x}')$, to previously selected conditions. The experimentalist labeled *least confident* chooses experiment conditions for which the theorist's best model is most uncertain,

[1] We ensured that logistic regression could qualitatively reproduce basic effects produced by each ground-truth agent when trained on the full data set.

i.e., for which the highest predicted outcome probability is closest to 0.5. The *model comparison* experimentalist identifies two candidate models from the logistic regression—one with all interaction terms (the candidate model) and one without interaction terms (the reference model)—and then selects experiment conditions for which the predictions of the two models disagree the most. Finally, the *falsification* experimentalist searches for experiment conditions under which the loss $\hat{\mathcal{L}}(M, X', Y', \vec{x})$ of the best candidate model is predicted to be the highest. This loss is approximated with a multi-layer perceptron, which is trained to predict the loss of a candidate model, M, given experiment conditions X' and dependent measures Y' that have already been probed.

Ground-Truth Agents

We evaluate each experimentalist across four experimental domains: object categorization, value-based decision making, the Stroop task, and task switching, each represented by a prominent computational model of human cognition (see Table 2). The corresponding computational models act as ground-truth agents and are used to generate observations y from experiment factors $\vec{x}$. We implemented each agent so that it outputs a probability (e.g., representing the likelihood of the agent choosing one response over another), such that the data produced by all agents are amenable to logistic regression.

Table 2: Ground Truth Agents. Agents correspond to quantitative models adapted from respective references and varied by the number of parameters as well as experiment factors.

Ground Truth	Reference	# Param.	# Fact.
Object Categ. Agent	Luce (1963)	1	4
Decision Making Agent	Tversky and Kahneman (1992)	6	4
Stroop Agent	Cohen et al. (1990)	25	6
Task Switching Agent	Yeung & Monsell (2003)	6	3

Object Categorization Agent Object categorization tasks require participants to categorize a target object (e.g., categorize a digit according to its parity), possibly while ignoring a distractor object (e.g., a second digit displayed below the target digit). The adapted Shepard-Luce choice rule from Logan and Gordon (2001) posits that the likelihood of an individual assigning a target object, represented as x, to a specific response category, represented as i, is proportional to their psychological similarity $\eta(x,i)$,

$$p(\text{“}x \text{ is } i\text{”}) = \frac{\eta(x,i)\beta_x + \varepsilon}{\sum_{j \in R} \eta(x,j)\beta_x + \sum_{j \in R} \eta(y,j)(1-\beta_x)} \quad (1)$$

where R corresponds to the set of all possible response categories and y represents a second object on display. Here, we assume an attentional bias toward processing the target object $\beta_x = 0.8$. Furthermore, we consider a scenario with only two response categories, i and j, with an equal attentional bias to each category. Finally, we assume processing noise $\varepsilon \sim \mathcal{N}(0, 0.01^2)$ with the constraint that $0 \leq p(\text{“}x \text{ is } i\text{”}) \leq 1$.

Experimental factors include the psychological similarities[2] between the target object and the two response categories $\eta(x,i)$, $\eta(x,j) \in \{1.25k \mid k \in \{1,2,\ldots,8\}\}$, as well as the similarities between the distractor object and the same two response categories $\eta(y,i)$, $\eta(y,j) \in \{1.25k \mid k \in \{1,2,\ldots,8\}\}$. The dependent measure is the probability of assigning category i to the target object x, $p(\text{“}x \text{ is } i\text{“})$.

Decision Making Agent Value-based decision making is concerned with how humans choose between offers, e.g., between two lotteries that are associated with different values and respective success probabilities. Prospect theory (Tversky and Kahneman, 1992) suggests that humans choose between two options, A and B, based on their expected values,

$$P(\text{Choose A}) = \frac{e^{1/T \cdot U(A)}}{e^{1/T \cdot U(A)} + e^{1/T \cdot U(B)}} \quad (2)$$

where $T = 0.1$ is the temperature of the choice softmax function and $U(A)$ and $U(B)$ are the expected utilities of the two choices. Here, we assume that each choice is associated with one value and a respective probability such that the expected utility of any option amounts to $U(A) = v(x_a)\pi(p_a) + \varepsilon$, where $v(x_a)$ corresponds to the subjective value of the outcome associated with option A and $\pi(p_a)$ corresponds to the probability of that outcome. We also assume that the computation is noisy, $\varepsilon \sim \mathcal{N}(0, 0.01^2)$. Following Tversky and Kahneman, 1992, we adopt an asymmetric value function,

$$v(x) = \begin{cases} x^{\alpha} & x \geq 0 \\ -\lambda(-x)^{\beta} & x < 0 \end{cases} \quad (3)$$

where $\alpha = \beta = 0.88$ and $\lambda = 2.25$. Furthermore, we assume that humans overweight low probabilities and underweight high probabilities,

$$\pi(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}} \quad (4)$$

where $\gamma = 0.69$ if $x < 0$ and $\gamma = 1.0$ if $x \geq 0$. The four experiment factors for this ground-truth agent correspond to the outcomes of both options, $x_a, x_B \in \{0.25k \mid k \in \{-4, 3, \ldots, 4\}\}$, and the respective probabilities, $p_a, p_b \in \{0.125k \mid k \in \{0, 1, \ldots, 10\}\}$. Finally, we considered $P(\text{Choose A})$ as the dependent variable.

Stroop Agent The Stroop Model is a neural network model that aims to explain cognitive processes involved in the ability of individuals to override habitual responses (e.g., reading

[2] Here, we assume that psychological similarities between objects and response categories are known prior to the experiment and, thereby, manipulable.

a color word "GREEN") in order to fulfill current task goals (e.g., naming the ink color of the color word). As described in Cohen et al. (1990), the model is comprised of two input layers, one representing the stimulus and one representing the task. The stimulus layer is divided into two groups, with each group consisting of two units. One group represents the color of the stimulus (e.g., green or red), and the other group represents the word (e.g., "GREEN" or "RED"). The two units constituting the task layer represent the color naming task and the word reading task. Both the stimulus and task inputs are multiplied by a matrix of connection weights, projecting from the input layers to an associative (hidden) layer. The resulting pattern of activity over the units in the associative layer is then passed through a logistic function, which is used to determine, through another set of connection weights, the pattern of activity over the output layer via a softmax function. The two units constituting the output layer represent the probabilities of the verbal responses, "green" and "red".

A fundamental assumption of the Stroop model is that the connections projecting from the word input units to the output layer (via the associative layer) are stronger than those of the color input units. This is believed to establish an automatic processing pathway for the word inputs. However, the model also proposes that the color input units may be selected over the word inputs (via projections from the task layer) at the associative layer. In this study, we applied the connection weights from Cohen et al., 1990 and added noise $\varepsilon \sim \mathcal{N}(0, 0.01^2)$ to the net input of each output unit.

The inputs to the Stroop Model can be represented as six experimental factors: the saturations of the two colors green and red $c_r, c_g \in \{0.1k \mid k \in \{0, 1, \ldots, 10\}\}$, the visibilities of the two words, "GREEN" and "RED" $w_r, w_g \in \{0.1k \mid k \in \{0, 1, \ldots, 10\}\}$, and the presence of the two tasks, color naming and word reading $t_c, t_w \in \{0, 1\}$. However, the experimental design space is constructed such that only one of the respective features and tasks is present, i.e., $c_i \geq 0, c_{j \neq i} = 0; w_i \geq 0, w_{j \neq i} = 0; t_i = 1, t_{j \neq i} = 0$. We treat the probability of responding "green" as the dependent measure.

Task Switching Agent One of the most robust findings in cognitive psychology is that individuals make fewer errors and respond more slowly when repeating a task as compared to switching from one task to another. Yeung and Monsell (2003) explain this and other task-switching phenomena with a simple model in which the activation level of a task is related to its performance,

$$P(\text{Correct Response to Task A}) = \frac{e^{1/T \cdot act_A}}{e^{1/T \cdot act_A} + e^{1/T \cdot act_B}}. \quad (5)$$

where $T = 0.2$ is the choice temperature, and act_A and act_B correspond to the activity of tasks A and B, respectively. The activity of any task is a non-linear function of its input, $act_i = 1 - e^{1.5 \cdot \text{input}_i}$. The input consists of multiple factors,

$$\text{input}_i = \text{strength}_i + \text{priming} \cdot \text{repetition} + \text{control}(i) + \varepsilon \quad (6)$$

including the degree of task practice, represented as strength$_i$, and a priming factor, which is incorporated when the task on the current experimental trial is identical to that of the previous trial (repetition $= 1$). Additionally, endogenous control is incorporated into the task input as a function of the task's strength (see Yeung and Monsell, 2003 for more details). Finally, noise $\varepsilon \sim \mathcal{N}(0, 0.01^2)$ is added to the input. For the simulations reported below, we parameterized the model according to Table 4 in Yeung and Monsell, 2003.

The three experimental factors of the task switching model are the strengths of the two tasks, strength$_i \in \{0.02k \mid k \in \{1, \ldots, 100\}\}$, and the type of task transition, repetition $\in \{0, 1\}$. We considered the probability of correctly responding to Task A as the dependent variable of interest.

Simulation Procedure

We evaluated each experimentalist by first generating an exhaustive "legal" set of experimental conditions (see the experiment factor bounds described for each ground-truth agent). Some experimentalists require seed data or models to determine novel experiment conditions. Thus, before initiating the first research cycle, we collected a seed data set that comprises ten evenly-spaced experiment conditions sampled from the legal set, along with the noisy observations obtained from the respective ground-truth agent. We also fitted a seed model to this data set. Together, the seed data set and the seed model provided the same initial condition for each experimentalist with a given ground-truth. Following this initialization, we ran 50 empirical research cycles, with each cycle consisting of the following steps: (1) sample ten novel experiment conditions according to the specific experimentalist strategy, (2) collect corresponding noisy observations from the ground-truth agent, (3) combine new data with existing data, and (4) fit a new candidate model to the data. For each cycle, we calculated the mean-squared error (MSE) of the theorist's candidate model's predictions for all experiment conditions (with noise-less observations). The entire procedure was independently simulated 20 times for each pairing of experimentalist and ground-truth agent, resulting in a total of 400 simulations.

Analysis

Prior to all analysis, we removed outliers, operationalized as an MSE greater or less than three standard deviations from the mean MSE for a given research cycle and experimentalist. We evaluated the performance for each pairing of experimentalist and ground truth by regressing the final MSE (after 50 cycles of the empirical research process) against the experiment strategy (categorical regressor). In alignment with our hypothesis, we contrasted each experimentalist strategy against the random experimentalist (intercept). Different seeds of the empirical research cycle were treated as random effects. To further examine the different experimentalist strategies, we calculated the variance across all collected observations from the dependent measure at the end of each simulation. As with the MSE, we regressed the variance against the experiment strategy (categorical regressor), with the random experimentalist as the baseline, and treated the seed for each cycle as a random effect. Finally, we projected the experiment conditions collected after 10 cycles[3] for each experimentalist onto a 2-dimensional plane using t-SNE.

Results

Table 3 and Figure 2 summarize the results for all pairings of experimentalist and ground truth. For the object categorization agent, we observe that none of the experimentalists yield a significantly lower MSE than the random experimentalist. Curiously, the model-driven experimentalists relying on falsification and model comparison strategies performed significantly worse than random sampling. For the value-based decision making agent, however, the falsification experimentalist yielded a significantly lower MSE compared to the random experimentalist. All other sampling strategies performed as well or worse as the random sampling strategy. For the Stroop agent, novelty sampling appeared as good as random sampling, whereas all other sampling strategies yielded a higher MSE compared to the random strategy. Finally, for the task switching agent, none of the sampling strategies yielded a lower MSE compared to the random sampling strategy. The model comparison and falsification strategies even yielded greater MSEs with each data collection cycle, resulting in significantly greater final MSEs as depicted in Figure 2D.

The right part of Table 3 indicates that most experimentalists yield lower variance across the collected observations when compared to the random sampling strategy, except, in some cases, for the novelty experimentalist. The same result was reflected in the projection of observed experiment conditions, grouped by sampling strategy (lower panels in Figure 2). In general, the model-driven experimentalists (falsification and model comparison) sample a narrow space of experiment conditions, which reflects a sampling bias. Conversely, data points collected by the random and novelty experimentalist appear more spread out, enabling them to capture the "bigger picture" of the ground truth. A post-hoc analysis revealed a strong correlation between the final MSE and variance of collected observations across seeds and experimentalists, r(395) = .51, p< .001, suggesting that the greater variance obtained by random sampling contributes to its success.

Discussion

Central bottlenecks in empirical research call for an integration of artificial intelligence into the scientific process. While solutions for automating theory discovery are emerging, there are few well-informed attempts at automating strategies for sampling experimental conditions in established psychological paradigms. In this article, we evaluated different experimental sampling strategies based on their abilities to effectively recover ground-truth agents from noisy observations. We found that the best sampling strategy depended on the ground truth to be discovered. Curiously, model-driven strategies more commonly underperformed random sampling (Table 4). Our results suggest that the relative disadvantage of the model-driven sampling of experiments results from the inability to capture variance across the entire design space, potentially due to strong sampling biases imposed by the model.

Table 3: Analysis Results. For every ground truth, experimentalists are compared based on the MSE of the resulting model, as well as the variance across all collected observations. The random sampling strategy acts as a reference level. Green/red highlighting indicates whether the respective experimentalist yields relatively lower/higher MSE or higher/lower variance, compared to random sampling.

Dependent Measure		MSE			Variance of Collected Data	
	β	S.E	p	β	S.E	p
Object Categorization						
Intercept (Random)	.0739	.0081	**<.001**	.0344	.0025	**<.001**
Novelty	.0042	.0052	.4173	-.0147	.0029	**<.001**
Least Confident	-.0023	.0052	.6579	-.0318	.0028	**<.001**
Model Comparison	.0287	.0052	**<.001**	-.0315	.0028	**<.001**
Falsification	.0188	.0052	**<.001**	-.0227	.0028	**<.001**
Decision Making						
Intercept (Random)	.3947	.0015	**<.001**	.1959	.0028	**<.001**
Novelty	-.0009	.0020	.6685	-.0071	.0039	.0733
Least Confident	.0008	.0020	.7086	-.1811	.0039	**<.001**
Model Comparison	.0051	.0020	**<.05**	-.1796	.0039	**<.001**
Falsification	-.0090	.0020	**<.001**	-.1655	.0039	**<.001**
Stroop						
Intercept (Random)	.2310	.0014	**<.001**	.1195	.0050	**<.001**
Novelty	-.0018	.0019	.3607	.0116	.0071	.1030
Least Confident	.0067	.0019	**<.001**	-.1152	.0071	**<.001**
Model Comparison	.0087	.0019	**<.001**	-.0854	.0071	**<.001**
Falsification	.0087	.0020	**<.001**	-.0707	.0071	**<.001**
Task Switching						
Intercept (Random)	.0467	0.0012	**<.001**	.0220	.0007	**<.001**
Novelty	-.0004	.0016	.7970	.0031	.0011	**<.01**
Least Confident	-.0020	.0016	0.2220	-.0196	.0011	**<.001**
Model Comparison	.0172	.0016	**<.001**	-.0193	.0011	**<.001**
Falsification	.0107	.0016	**<.001**	-.0121	.0011	**<.001**

Our results comport with prior simulations of Dubova et al. (2022), demonstrating that random sampling of experiment conditions can outperform active sampling stategies. The simulation approach taken by Dubova et al. (2022) differed in two important ways: (a) ground truths were represented by multi-variate Gaussian distributions instead of established models of human cognition, and (b) scientific models were approximated by autoencoders instead of more interpretable and more commonly applied logistic regression models. Yet, results from both approaches converge in that they identify advantages of random sampling over model-driven experimentation. Contrary to Dubova et al. (2022), our study sug-

[3] We chose a small number of cycles to avoid obscuring differences across experimentalists due to visual cluttering.

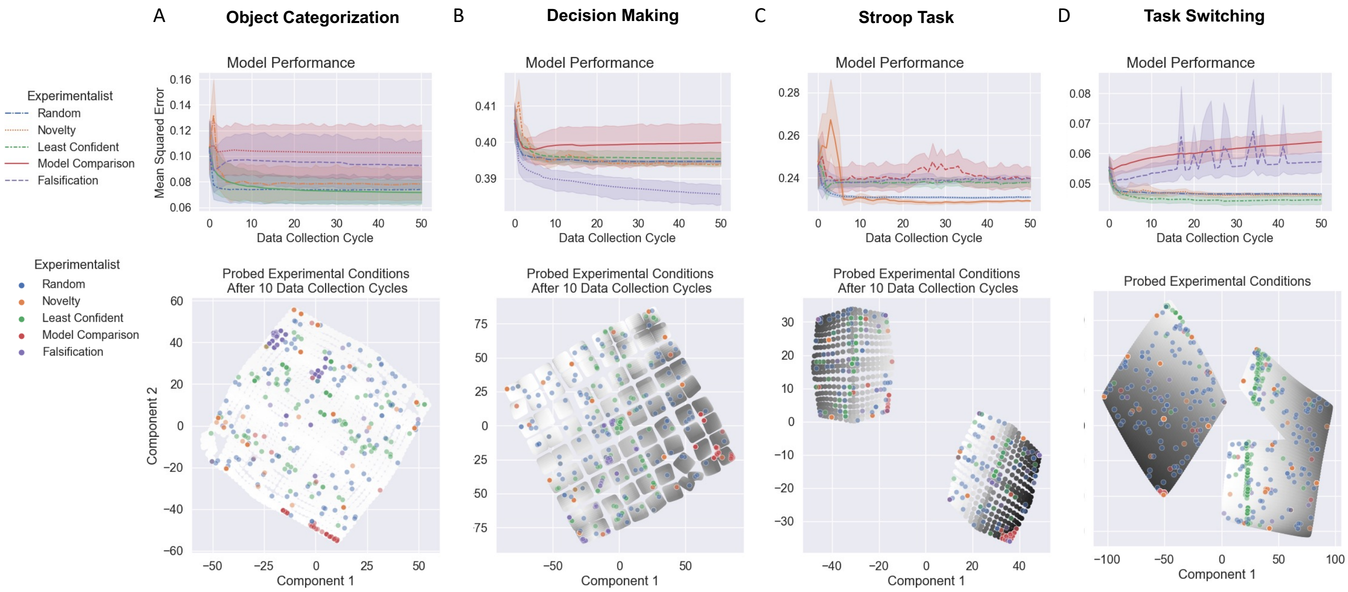

Figure 2: Simulation Results. Top panels show the error of the theorist model across the full set of legal experimental conditions for the (A) object categorization agent, (B) value-based decision making agent, (C) Stroop agent, and (D) task switching agent. Different colors represent different experimentalists. Bottom panels depict a t-SNE projection of all experimental conditions (sampled after 10 data collection cycles) for the respective ground-truth agent onto a two-dimensional plane. The exhaustive space of legal experimental conditions is represented by white-gray points, with the degree of grayness indicating the value of the dependent measure for the respective experimental condition. Darker shades of gray correspond to higher probabilities. Colored points represent the final set of experimental conditions that were sampled by the best experimentalist of each type.

Table 4: Summary. The number of times that experimentalists performed better, worse, or equal to random sampling.

Experimentalist	Better	Worse	Equal
Novelty	0	0	4
Least Confident	0	1	3
Model Comparison	0	4	0
Falsification	1	3	0

gests that, active experimentation can, in principle, outperform random sampling. However, if the ground truth is unknown, random sampling can be a successful heuristic compared to most other approaches. This is consistent with observations that matrices with low rank (e.g., structured data from experiments with few independent variables) are well approximated with random sampling (Halko et al., 2011).

While suggestive, the evidence provided by this study is limited by its specific parameterization of the empirical research cycle. First, we benchmarked different experimentalists on only a small subset of quantiative models that served as ground truth. As our analyses indicate, the best experimentation strategy may depend on the particularities of the phenomenon under investigation. Consequently, future work should examine the performance of different strategies as a function of features characterizing the object of study, such as the rank of the data matrix (Halko et al., 2011). Second, we restricted model discovery to fitting parameters in a logistic regression. While this yields interpretable models, it does not involve the most challenging aspect of theory discovery—identification of the model architecture. Autonomous theorists could deploy different forms of automated model discovery, such as symbolic regression (Guimerà et al., 2020; Schmidt and Lipson, 2009; Udrescu et al., 2020) or neural architecture search (Elsken et al., 2019; Musslick, 2021). Finally, we used only simple experimentalists strategies. In principle, we could compose or serialize strategies into more sophisticated experimentalists and counterbalancing schemes (Musslick et al., 2022). The latter may involve eliminating or adding independent variables to explore boundaries of the experimental design space (Dubova et al., in press). To address these limitations, we open-sourced a framework for automating and simulating steps of the empirical research process (`https://github.com/autoresearch/autora`), which offers a principled way to examine the differential effects of ground-truth agents, strategies for theory discovery, and methods for experiment sampling, on scientific discovery.

In conclusion, the work presented in this study suggests that random sampling of experimental conditions is a valuable heuristic compared to model-driven experimentation. Furthermore, while the prospect of large-scale experimentation and automated scientific discovery seems appealing, our work highlights the need for more comprehensive metascientific studies to identify effective strategies for sampling experiments from large design spaces. We hope that the metascientific test bed introduced in this study will encourage a proliferation of efforts along these lines and pave the way for autonomous empirical research.

Acknowledgements

This research was supported by Schmidt Science Fellows, in partnership with the Rhodes Trust, as well as the Carney BRAINSTORM program at Brown University.

References

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, *117*(16), 8825–8835.

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2022). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, 1–55.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the stroop effect. *Psychological review*, *97*(3), 332–361.

Dubova, M., Sloman, S. J., Andrew, B., Nassar, M. R., & Musslick, S. (in press). Explore your experimental designs and theories before you exploit them! *Behavioral and Brain Sciences*.

Dubova, M., Moskvichev, A., & Zollman, K. (2022). Against theory-motivated experimentation in science. *MetaArXiv preprint: https://doi.org/10.31222/osf.io/ysv2u*.

Elsken, T., Metzen, J. H., Hutter, F., et al. (2019). Neural architecture search: A survey. *JMLR*, *20*(55), 1–21.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Guimerà, R., Reichardt, I., Aguilar-Mogas, A., Massucci, F. A., Miranda, M., Pallarès, J., & Sales-Pardo, M. (2020). A bayesian machine scientist to aid in the solution of challenging scientific problems. *Science Advances*, *6*(5), eaav6971.

Halko, N., Martinsson, P.-G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, *53*(2), 217–288.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., et al. (2009). The automation of science. *Science*, *324*(5923), 85–89.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1993). Dendral: A case study of the first expert system for scientific hypothesis formation. *Artificial intelligence*, *61*(2), 209–261.

Logan, G. D., & Gordon, R. D. (2001). Executive control of visual attention in dual-task situations. *Psychological review*, *108*(2), 393–434.

Luce, R. D. (1963). Detection and recognition.

Musslick, S. (2021). Recovering quantitative models of human information processing with differentiable architecture search. In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society* (pp. 348–354).

Musslick, S., Cherkaev, A., Draut, B., Butt, A. S., Darragh, P., Srikumar, V., Flatt, M., & Cohen, J. D. (2022). Sweetpea: A standard language for factorial experimental design. *Behavior Research Methods*, *54*(2), 805–829.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *science*, *324*(5923), 81–85.

Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report No. 1648). University of Wisconsin–Madison.

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of experimental psychology*, *55*(6), 509–523.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297–323.

Udrescu, S.-M., Tan, A., Feng, J., Neto, O., Wu, T., & Tegmark, M. (2020). AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *arXiv preprint arXiv:2006.10782*.

Yeung, N., & Monsell, S. (2003). Switching between tasks of unequal familiarity: The role of stimulus-attribute and response-set selection. *Journal of Experimental Psychology: Human perception and performance*, *29*(2), 455–469.