Against theory-motivated experimentation in science

Marina Dubova[1,2,3], Arseny Moskvichev[4], Kevin Zollman[2,3]

[1]Cognitive Science Program, Indiana University, Bloomington, IN USA

[2]Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA USA

[3]Center for Formal Epistemology, Carnegie Mellon University, Pittsburgh, PA USA

[4]Department of Cognitive Sciences, University of California, Irvine, CA USA

Contact: marina.dubova.97@gmail.com

Abstract

Scientists must choose which among many experiments to perform. We study the epistemic success of experimental choice strategies proposed by philosophers of science or executed by scientists themselves. We develop a multi-agent model of the scientific process that jointly formalizes its core aspects: active experimentation, theorizing, and social learning. We find that agents who choose new experiments at random develop the most accurate theories of the world. The agents aiming to confirm, falsify theories, or resolve theoretical disagreements end up with an illusion of epistemic success: they develop promising accounts for the data they collected, while completely misrepresenting the ground truth that they intended to learn about. Agents experimenting in theory-motivated ways acquire less diverse or less representative samples from the ground truth that are also easier to account for. The novelty-seeking agents suffer from collecting unrepresentative observations. Thus, random data collection combines virtues of diverse and representative sampling from a target scientific domain which enables cumulative development of the successful theoretical accounts of it. We suggest that randomization, already a gold standard within experiments, is also beneficial at the level of experiments themselves.

Significance Statement

While it is natural to think that targeted, theory-driven, experiments — based on the "current wisdom" accumulated by scientists so far — are the best way to improve knowledge, our results show that far more unstructured (random) investigations are robustly preferred across a wide range of conditions. We develop a new model of scientific process, which is the first joint formalization of active data collection, theorizing, and social learning. We formally investigate epistemic success of the theory-driven experimentation strategies, such as verification, falsification, and crucial experimentation. Some of these strategies have so far been considered gold standards across scientific fields. Our multi-agent simulations suggest that these strategies lead to an illusion of epistemic success, rather than better theories about the world.

1. Introduction

1.1. The problem of experimental choice

Every empirical scientist faces a question of how to choose the next experiment. Some would pick an experiment likely to produce data favoring their theory, some would attempt to falsify their theory or resolve major theoretical disagreements, while others would just test a relationship between variables of interest and look for the theoretical motivation post-hoc. Despite the crucial importance of experimentation choice for sciences, the methodological recommendations on this front are rather limited.

Several broad strategies for experimental choice can be found in the philosophical literature. Fleck and Kuhn famously argued that scientists usually conduct experiments where the outcomes are already well known, or at least broadly expected (Fleck, 1979; Kuhn, 1970 – whether this is what they thought we ought to do is left for another discussion; also, see Klayman & Ha, 1987). Popper suggested that scientists should seek out those experiments most likely to falsify their theories (Popper, 1963), and in a related vein Mayo has suggested that scientists should subject their theories to severe tests (Mayo, 2018). Lakatos, and others, highlighted the importance of crucial experiments which distinguished between one theory and another (Lakatos, 1974). These alternatives share a crucial assumption that theory-motivated experimentation, in one way or another, facilitates theory-building and learning about the world. This stands in contrast to contemporary discussions in statistics and machine learning which highlight the importance of unbiased, random sampling in many circumstances (e.g. Baribault et al., 2018; Sloman et al., 2022).

Even though choice of experiment constitutes an essential everyday part of scientific activity, the discussions over the efficiency of different strategies remain largely unresolved. Scientists who engage in experimentation on an everyday basis are left to their own intuitions and social conventions of their fields. Psychologists, for example, primarily conduct new experiments to corroborate or (sometimes) challenge the century-old theories, tend to use

similar paradigms and experimental settings, and despise experiments that are not informed by a developed theory (see van Rooij & Baggio, 2021 for a recent proposal). Geneticists, on the other hand, often explore the relationships over vast amounts of genes and other variables, without any theory-driven hypothesis of what the data can show (e.g. Palmer et al., 2022).

1.2. Formal assessment of data collection strategies: does theory-motivated data collection help to develop better theories?

We develop a general modeling framework to study the epistemic success of different strategies that communities of interacting scientists may follow when trying to learn about the world. In this multi-agent model, agents collect data, try to explain them, and communicate their findings and theoretical accounts to each other (Figure 1). This framework enables us to investigate essential aspects of scientific activity – experimentation, explanation, and communication – as well as their possible interactions (e.g. one's theoretical framework influencing the data one collects). We apply the model to computationally probe one of the fundamental intuitions of the broad scientific community about experimentation: the primacy of theory-driven experimentation.

Here, we rigorously assess the epistemic success of different experimentation strategies in different contexts. We formalize the commonly proposed experimentation strategies: confirmation (Fleck, 1979; Kuhn, 1970), falsification (Popper, 1963), crucial experimentation (Lakatos, 1974), novelty-driven experimentation (in a sampling literature: Shi & Tang, 2021), random baseline (Brunswik, 1955), and their hybrids. To achieve the most robust and general results, we extensively vary other potentially important components, such as complexity of the "ground truth" that the agents are learning about, the ways in which they interact, their meta-theoretic preferences, measurement resources, and other factors.

Across all the conditions, we found that the agents collecting data at random end up with the best theoretical accounts for the ground truth. The agents following confirmation-,

falsification-, theoretical disagreement-based strategies and their hybrids limited their observations to a simpler subset of the ground truth that did not let them produce a successful theory about it, but created an illusion of such. We therefore conclude that theoretically motivated experiment choice is potentially damaging for science, but in a way that will not be apparent to the scientists themselves.

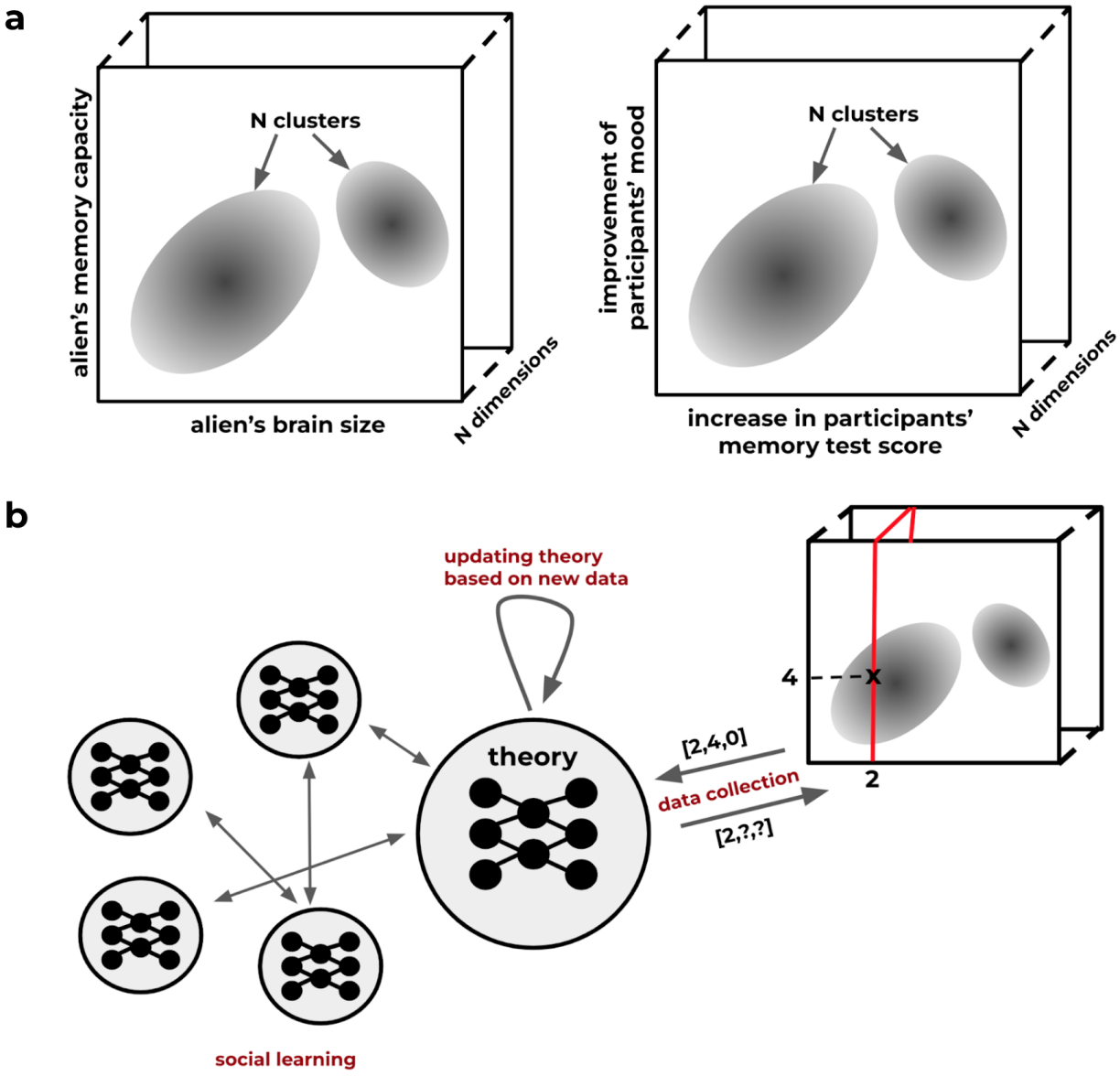

Figure 1. **a**, Illustration of the ground truth and its' example interpretations. The ground truth is represented as a mixture of multivariate gaussian distributions with N clusters that span across D dimensions (see 4.1.1.). Modeling truth this way provides an idealized representation of

several aspects of the scientific process. On one level of abstraction, the ground truth and its dimensions can be interpreted as reflecting the distribution of different measurable properties of one's target phenomena (left). For example, an alien species will have multiple properties reflected by scientists' available measurements: height, weight, results on memory test, brain size, social group size, and so on. Some of these properties are correlated (e.g. height and weight), but some are independent. Some dimensions follow a nonlinear relationship (e.g. here: memory capacity and brain size). The dimensions do not have to reflect the distribution of properties, however: they can also be interpreted as scientific manipulations and their outcomes (right). For example, in a given study a scientist can manipulate participants' mood, to a particular extent, and compare their memory test scores before and after this manipulation. The results of a family of such experiments can be distributed in many ways. Figure illustrates one potential scenario with two kinds of cognitive processes that react to scientists' mood manipulation differently: one produces a steady improvement in participants' test scores, the other process reacts to the mood change in a reverse manner, but responds to a smaller range of participant's mood manipulations in the experiment.
b, Illustration of the model. The agents (scientists) in a group are trying to develop compressed descriptions of the multi-dimensional ground truth (theories; formalized as simple autoencoders with one hidden layer). The agents collect the data by controlling a value along one dimension of the ground truth (see 4.1.3.); the values along other dimensions of an observation are conditionally sampled from the ground truth. The agents update their theories (see 4.1.2.) whenever they obtain a new observation. The agents interact with other agents in a group, by sharing their observations or theories (see 4.1.4.).

2. Results

We conducted three simulation experiments to investigate a role of experimentation strategy (Table 1) on the epistemic success of the learners. First, we looked at the main effects of the agents' experimentation strategy on their "subjective" and "objective" epistemic success across contexts (see 2.1.). We found that random experimentation leads agents to develop the best accounts for the ground truth in nearly all the contexts we considered. Agents collecting data in a theoretically informed way, even if guided by falsification or disagreement, ended up sampling less diverse or less representative observations from the ground truth, which are easier to account for. These agents developed theories that appeared very successful, but in fact misrepresented the ground truth. Then, we tested whether theoretically informed

experimentation can be justified when the agents are limited in the amount of observations they can collect (see 2.2.) or when they have good theories (see 2.3.). We found that neither limited learning time nor well-developed theories help agents to develop better theories than when they simply choose experiments at random.

strategy	intuition
confirmation	the agent samples a new observation close to the previously collected observation (referent) that is well explained by its theory
falsification	the agent samples a new observation close to the previously collected observation that is poorly explained by its theory
disagreement	the agent samples a new observation close to the previously collected observation on which predictions of its and another agent's theories largely disagree
disagreement + confirmation	the agent samples a new observation close to the previously collected observation that is both well explained by its theory and also leads to a different prediction by another agent's theory
disagreement + falsification	the agent samples a new observation close to the previously collected observation that is both poorly explained by its theory and also leads to a largely divergent prediction by another agent's theory
novelty	the agent samples a new observation that is very different from its previously collected observations
random	the agent samples a new observation at random

Table 1. Intuition behind the tested experimentation strategies (see 4.1.3.). Theory-motivated strategies are probabilistic, so that the referents for sampling are chosen with respect to their confirmation-, falsification-, or disagreement- score (or both scores, in hybrid strategies). Note that the referent in theory-motivated experimentation only determines a value along one dimension of a future observation (controlled dimension), while all other observations are sampled from the ground truth distribution conditioned on that value. Agents following all strategies except random start with 10 randomly collected observations, and then have a 10% random exploration rate.

2.1. Epistemic success of experimentation strategies across contexts

First, we simulated the model to determine the relative success of experimentation strategies across contexts: for different complexities of the ground truth, agents' communication strategies, agents' preference for simpler explanations, measurement capacities, group sizes, and others. Multi-agent learning with each combination of parameters determining the "context" (see Methods) was simulated 3 times, resulting in 9060[1] simulations analyzed further.

2.1.1. Subjective performance

First, we looked at the influence of experimentation strategy on how well agents' resulting theories accounted for only the data they actually collected (subjective success), thus measuring agents' perceived epistemic success. In our analysis of subjective performance, we did not consider how the agent's theories would account for data that was not collected by any agent in the community or against the ground truth itself. For our main descriptive analysis of the simulation results, we applied linear regression with the formula:

per-group average subjective theoretical error ~ experimentation strategy + N of agents + measurement capacity + explanation capacity + N of clusters in the ground truth + N of dimensions in the ground truth + collective interaction strategy (Model 1)

to estimate the effects of data sampling strategies on agents' subjective performance, while controlling for the effects of all other variables we varied between simulations. We replicated the analysis with regression models that include an interaction of the experimentation strategy with the group communication strategy or with the ground truth complexity (number of dimensions and clusters in the ground truth) to estimate how context-dependent the main effects of different experimentation strategies are. We also replicated the analyses using the

[1] Several simulations (2) did not finish in time, which is controlled for in the analysis.

group's best individual's subjective reconstruction error, rather than the group average, as a target variable; and with respect to the subjective scores of agents' at different points of their learning progress. We performed post-hoc contrast tests to construct the ranking of the data sampling strategies according to their subjective success, based on the results from Model 1 (Table 2).

Across the analyses, we found that the agents following random or novelty-based experimentation strategies ended up with worse accounts for their collected observations than the agents following all other experimentation strategies that we tested (Figure 2a). This accounts for the "illusion" of success that we discussed in the introduction: the agents appear to themselves to have successfully captured the truth. These results replicated with respect to both average subjective success of the group and its best individual's subjective score. The ranking of the strategies with respect to their subjective success remained the same after possibilities of the interactions were added, indicating that the main effects are quite stable across the learning contexts. These results were also stable across time, resulting in comparable rankings of the strategies based on the groups' subjective performances measured after 50, 100, 150, 200, 250, and 300 observations were collected by the group.

In sum, the agents following novelty-based and random experimentation strategies generated less successful accounts for their own observations than agents following all other experimentation strategies, no matter how complex the "ground truth" was, how the agents communicated about their results, how many agents were learning together, how elaborate theories they were building, how limited they were in their measurements, and how many observations they collected.

2.1.2. Objective performance

We performed the same analyses to assess the influence of data sampling strategies on agents' objective performance, evaluating them against the representative samples from the full

"ground truth" distribution instead of their own collected datasets. This represents the actual success of scientific theories in accounting for the truth, not the apparent success captured by our subjective performance measure.

Here, the relative success of different data sampling strategies mostly reversed. The agents collecting data at random ended up capturing the ground truth better than agents following all other strategies. Confirmation-based, falsification-based, disagreement-based strategies, and their hybrids performed significantly worse than the random strategy, while the "novelty-based" experimentation performed either on par or worse than the random strategy, depending on the type of analysis (Table 2, Figure 2a). The superiority of the random sampling strategy replicated across all our analyses: when the model included only the main effects, interaction between the data sampling strategy and the group communication strategy, and interaction between the data sampling strategy and the properties of the ground truth (dimensionality and number of clusters).

Then, we looked at the relative objective epistemic success of the data sampling strategies depending on the number of observations that the agents have collected from their environment. Again, we found that the random strategy leads the agents to come up with either the best or same-quality accounts for the ground truth as all other strategies, measured after the agents have learned from 50, 100, 150, 200, 250, and 300 samples. Thus, random experimentation outperforms all other tested strategies across a variety of parameters, including the type of ground truth, number of observations, measurement limitations, communication strategy, learning time, and others.

Subjective average performance	**Subjective best score**	**Objective average performance**	**Objective best score**
confirmation	confirmation	random	random
disagreement + confirmation	disagreement + confirmation	novelty	novelty
disagreement	disagreement	disagreement + falsification	disagreement + falsification
disagreement + falsification	disagreement + falsification	falsification	falsification
falsification	falsification	disagreement	disagreement
random	random	disagreement + confirmation	disagreement + confirmation
novelty	novelty	confirmation	confirmation

Table 2. Rankings of the strategies with respect to their subjective & objective average performance, subjective & objective best individual score (ranked from the most successful on the top to the least successful on the bottom). The solid lines between the strategies represent significant ($p<0.05$) post-hoc differences between the strategy and all other strategies across the line (Tukey test). The dotted line denotes that the strategy is significantly different from only some of the strategies across the line.

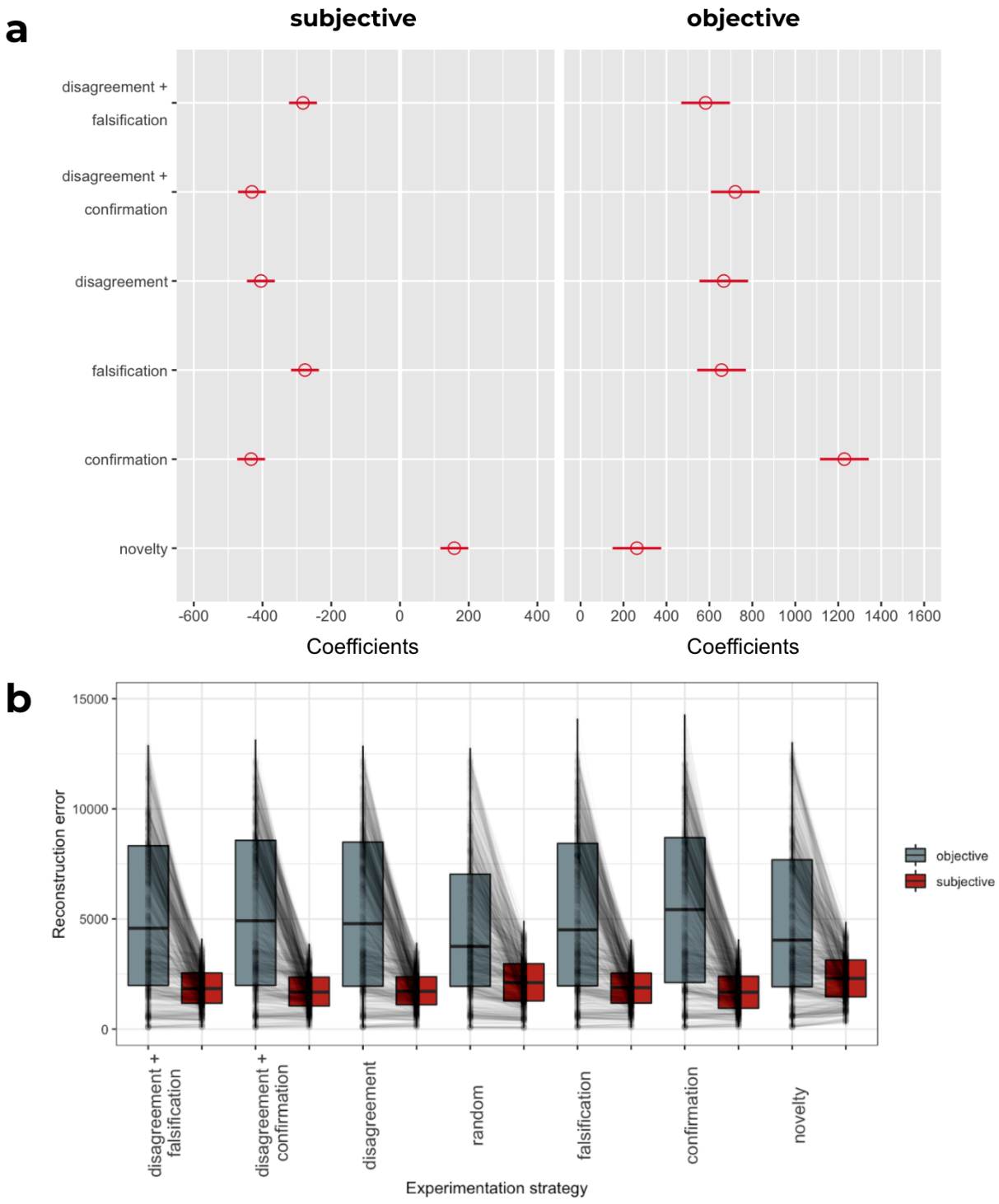

Figure 2. **a,** Subjective and objective theoretical error: estimates of experimentation strategies' effects in Model 1 (model fit for subjective performance analysis: $F(19,9040) = 1116.02$, $p = 0.000$ $R^2 = 0.70$; model fit for objective performance analysis: $F(19,9040) = 2572.182$, $p = 0.000$; $R^2 = 0.844$). All the coefficients are obtained relative to the baseline random strategy: positive

coefficients indicate that the theoretical (reconstruction) error is higher than the baseline (random experimentation), negative coefficients indicate lower than the baseline (random experimentation) error. **Subjective performance analysis** (left): disagreement + falsification: {est=-282.26; 95% CI [-322.83, -241.68]; t=-13.64; p=0.000}; disagreement + confirmation: {est=-430.81; 95% CI [-471.38, -390.23]; t=-20.81, p=0.000}; disagreement: {est=-404.80; 95% CI [-445.37, -364.22]; t=-19.56; p=0.000}; falsification: {est=-276.42; 95% CI [-316.99,-235.85], t=-13.35, p=0.000}; confirmation: {est=-433.23; 95% CI [-473.80, -392.66], t=-20.93, p=0.000}; novelty: {est=158.14; 95% CI [117.57, 198.71]; t=7.64, p=0.000}. **Objective performance analysis** (right): disagreement + falsification: {est=582.24; 95% CI [468.874, 695.625]; t=10.067; p=0.000}; disagreement + confirmation: {est=720.45; 95% CI [607.07, 833.82]; t=12.45; p=0.000}; disagreement: {est=666.92; 95% CI [553.55, 780.30]; t=11.53; p=0.000}; falsification: {est=656.62; 95% CI: [543.25, 770.00]; t=11.35; p=0.000}; confirmation: {est=1228.46; 95% CI [1115.09, 1341.84]; t=21.24; p=0.000}; novelty: {est=262.37; 95% CI: [149.00, 375.75]; t=4.54; p=0.000}.
b, Relationship between subjective and objective reconstruction error. The average objective and subjective theoretical error scores for each simulation are connected with a line. Notice that the lines connecting objective and subjective performance for each simulation rarely intersect, thus, reflecting the positive relationship between them.

2.1.3. Subjective and objective epistemic success

One might be tempted to interpret the results so far as a simple case of overfitting. Agents who find tighter fits to their observed data will do worse than those with slightly worse fits that are predictively more accurate. In order to investigate this possibility, we analyzed the direct relationship between subjective and objective epistemic success of the agents.

The relationship between objective and subjective performance was significant and positive (Kendall's tau=0.35, z=50.39, p<0.001), indicating that a particular group of agents who performed better-than-average subjectively was more likely to perform better-than-average objectively as well. The positive significant effect replicated with the best objective and subjective scores in the group as the variables (Kendall's tau = 0.38, z=54.87, p<0.001). Therefore, while the expected objective and subjective epistemic successes for each experimentation strategy are almost inversely related (see 2.1.1., 2.1.2., Table 2, Figure 2a), the particular group's higher subjective performance is predictive of its higher objective performance (Figure 2b). For

example, if a group of agents following confirmation strategy is performing subjectively worse than another group following the confirmation strategy, we expect the first group to also be less successful objectively. This relationship almost reverses when the simulations are grouped according to the experimentation strategy: if the strategy seems more successful than another strategy according to subjective epistemic success, it will most likely be less successful objectively (Figure 2b). This result indicates that the group's subjective performance is generally aligned with the objective metrics of epistemic success, while the two are conflicting at the experimentation condition level.

2.1.4. Within-group theoretical heterogeneity

We analyzed the influences of data sampling strategy on within-group theoretical heterogeneity, sometimes known as cognitive diversity (Kitcher, 1993). To measure heterogeneity, we 1) computed the mean euclidean distance between the autoencoder weights (theory) of all pairs of agents in a group after aligning their internal nodes (*heterogeneity of theories*) and 2) asked all the agents in a group to generate predictions on 10000 random samples from the ground truth and computed the average Euclidean distance between such predictions (*heterogeneity of theory-based predictions*) (for more details, see 4.3.). According to both metrics, the groups following the confirmation strategy became more heterogeneous than the groups following other experimentation strategies. Moreover, the agents following all other strategies ended up with comparable within-group theoretical heterogeneity (Table 3).

Then, we looked at the relationship between a group's theoretical heterogeneity and its objective epistemic success. We used the regressions that included all the main effects (Model 1) and the theoretical heterogeneity scores as predictors. We found significant effects of both types of theoretical heterogeneity on agents' theoretical success. The more heterogeneous theoretical frameworks a group developed, the higher objective theoretical errors it ended up with (Group average theoretical error analysis; heterogeneity of theory-based predictions: {est=16.67; 95% CI

[15.89,17.46]; t=41.66; p=0.00}, heterogeneity of theories: {est=11.66; 95% CI [11.23,12.10]; t=52.88; p=0.00}. Group best individual's theoretical error analysis; heterogeneity of theory-based predictions: {est=6.39; 95% CI [5.99,6.79]; t=31.28; p=0.00}, heterogeneity of theories: {est=14.61; 95% CI [13.94,15.27]; t=43.06; p=0.00}). This result stands in contrast to other results that postulate an important relationship between cognitive diversity and theoretical success (Kitcher, 1993; Hong & Page, 2004; Zollman, 2010). It is important to note, however, that our model did not include some features present in these other models, and thus should not be taken as a refutation.

Heterogeneity of theory representations	**Heterogeneity of theory-based predictions**	**Individual sampling variability**	**Individual sampling variability over time**	**Between-agent sampling variability**	**Between-agent sampling variability over time**	**Representativeness of samples**	**Representativeness of samples over time**
confirmation	confirmation	novelty	novelty	novelty	novelty	confirmation	confirmation
novelty	disagreement + confirmation	random	random	random	falsification	random	random
disagreement + confirmation	random	disagreement + falsification	disagreement + falsification	falsification	disagreement + falsification	disagreement	disagreement + falsification
disagreement	falsification	falsification	falsification	disagreement + falsification	disagreement + confirmation	falsification	falsification
falsification	disagreement	disagreement	disagreement	disagreement	disagreement	disagreement + confirmation	disagreement
disagreement + falsification	novelty	disagreement + confirmation	disagreement + confirmation	disagreement + confirmation	random	disagreement + falsification	disagreement + confirmation
random	disagreement + falsification	confirmation	confirmation	confirmation	confirmation	novelty	novelty

Table 3. Rankings of the strategies with respect to their scores on the analysis metrics (higher to lower) based on the Tukey post-hoc comparisons of the linear model (Model 1) coefficients of the experimental strategies with each metric used as target variable. The solid lines between the strategies represent significant post-hoc differences between the strategy and all other strategies across the line (Tukey test). The dotted line denotes that the strategy is significantly different from only some of the strategies across the line.

2.1.5. Agents' sampling behavior

To further understand consequences of following different types of experimentation strategies and their potential relation to agents' epistemic success, we developed a number of metrics to analyze agents' sampling behavior. Namely, we looked at the between- and within-agent sampling diversity, representativeness of sampled observations, and the way these properties of sampling behavior change over the time of learning (for details, see 4.4.).

Agents' following theoretically-informed experimentation strategies collected less diverse samples from the ground truth than the agents collecting data in novelty-based or fully random ways. Moreover, the theory-informed experimentalists collected less and less diverse samples the more they learned. On the group-level, the agents' who followed the confirmation strategy collected similar data which were becoming more and more similar over time. The agents' driven by novelty, on the other hand, collected different observations which were becoming more distinct further in learning. Novelty-driven agents, however, suffered from collecting much less representative samples from the ground truth than agents following all other strategies (Table 3, Figure 3).

While we did not find an epistemic benefit to theoretical diversity, these results indicate that there is an indirect epistemic benefit to the collecting of diverse data through diverse experimentation (see also Zollman, 2010). Our novel result suggests that this diversity will appear counterproductive to the scientists themselves (subjective performance) while it is in fact indicative of more successful science (objective performance).

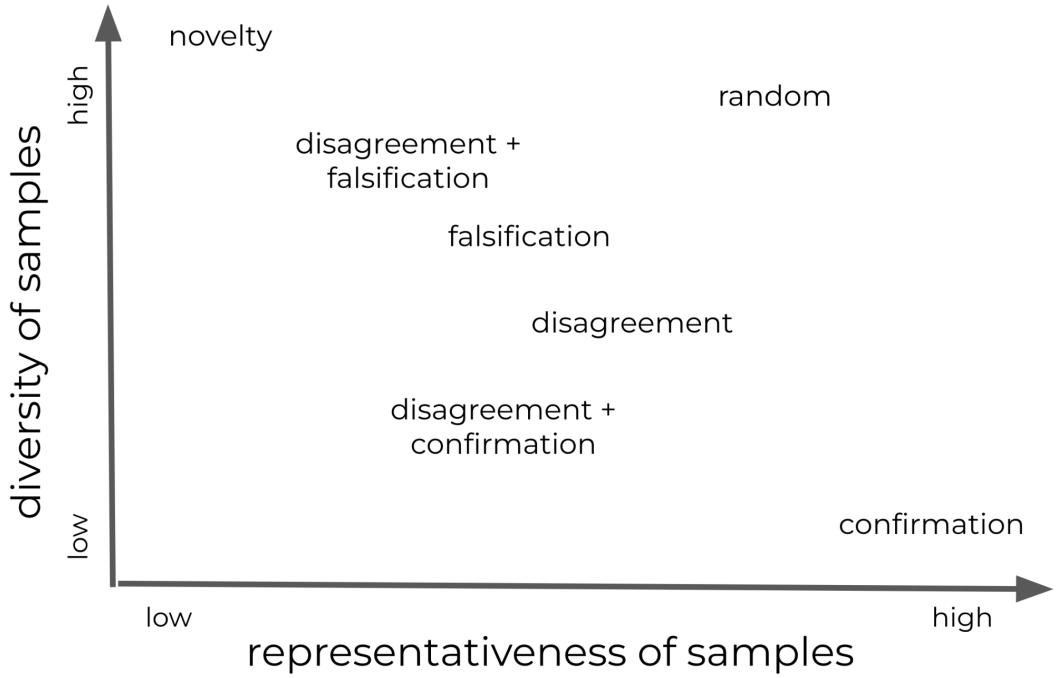

Figure 3. Mapping of the experimentation strategies with respect to average representativeness and diversity of samples that they lead to.

2.2. Learning with very limited observations

We hypothesized that the novelty-based strategy might be superior to the random strategy in the very beginning of learning, allowing the agents to get more varied and informative samples from the ground truth (as in Shi & Tang, 2021; Young, Cole & Sutherland, 2009). We performed new simulations, where the agents either conduct all their experiments at random or start to conduct experiments in a novelty-driven way after collecting only one (instead of 10, as in previous simulations) observation at random. In these simulations, we looked at the subjective and objective epistemic success of the agents with better temporal resolution, recording their scores each time the group collected 5 new empirical observations. All other parameters of the simulations were extensively varied as in the first experiment (see 2.1.), and every condition was simulated 5 times, resulting in 4191[2] simulations in total.

[2] Several simulations (129) did not finish in time, which is controlled for in the analysis.

In contrast to our expectations, we found that even in the very beginning of learning, the novelty-based experimentation strategy does not allow the agents to supersede the benefits of more representative, fully random experimentation. The superior objective performance of agents sampling the data at random is already recognizable after they collect 40 observations, and it only increases throughout learning (Figure 4).

Figure 4. Objective theoretical error: estimate of the effect of novelty-motivated vs. random (baseline) experimentation in Model 1, across learning time. Positive coefficients indicate that the theoretical error is higher than the baseline, negative coefficients indicate lower than the baseline reconstruction error. Notice that agents' following novelty vs. random strategy do not produce significantly different successful theories in the beginning of learning, but start diverging in their theoretical success with time. Agents collecting the data at random steadily increase their theoretical advantage with time. The difference between the strategies in the average objective performance that they lead to becomes significant after a group collects 60 observations (novelty: est=65.43, 95% CI [4.54,126.31], t=2.11, p=0.04) and monotonically increases afterwards. The difference in the best individual's objective score becomes significant after a group collects 40 observations (novelty: est=69.51, 95% CI [8.99,130.02], t=2.25, p=0.02) and then only monotonically increases.

2.3. "Good theories" do not save theory-informed experimentation

In all the simulations thus far, agents began with a random theory that they refined over time in response to collected data. We hypothesized that having substantial prior knowledge (good theories) about the ground truth might help the agents to benefit from theory-informed experimentation. We tested this intuition by pre-training each agent's theory on 10, 50, or 100 prior observations randomly sampled from the ground truth. Then, we let the agents learn with a predetermined experimentation strategy (as described in section 4.3.) and recorded their subjective and objective performance each time 5 new observations were collected in a group. Note that the groups pretrained with 100 observations per agent ended up collecting less observations with their target strategy than they were pre-trained with. Here, we varied all other parameters in the same way as in experiments 1 and 2, which resulted in 8906[3] simulations.

We analyzed the data with a linear model that included all contextual variables from the simulations (as in Model 1) as well as the amount of pretraining observations and its' interaction with agents' experimentation strategy as predictors. We found that the pre-training and its extent did not affect the relative epistemic success of the strategies (N of pretraining observations: {est=0.73; 95% CI [-3.22,4.68]; t=0.36; p=0.72}; N of pretraining observations × disagreement & falsification: {est=0.33; 95% CI [-5.26,5.92]; t=0.12; p=0.91}; N of pretraining observations × disagreement & confirmation: {est=-0.61; 95% CI [-6.20,4.98]; t=-0.21; p=0.83}; N of pretraining observations × disagreement: {est=-0.87; 95% CI [-6.45,4.72]; t=-0.30; p=0.76}; N of pretraining observations × falsification: {est=1.34; 95% CI [-4.24,6.93]; t=0.47; p=0.64}; N of pretraining observations × confirmation: {est=-1.91; 95% CI [-7.50,3.68]; t=-0.67; p=0.5}; N of pretraining observations × novelty: {est=-0.10; 95% CI [-5.69,5.49]; t=-0.04; p=0.97}). After starting to sample in an informed way, the agents' theories start to become less and less objectively successful than the theories constructed by agents that continue to sample fully at random. Overall, the relative ranking of strategies is quite stable over time and independent of the amount of prior knowledge (Figure 5).

[3] Several (166) simulations did not finish in time, which is controlled for in the analysis.

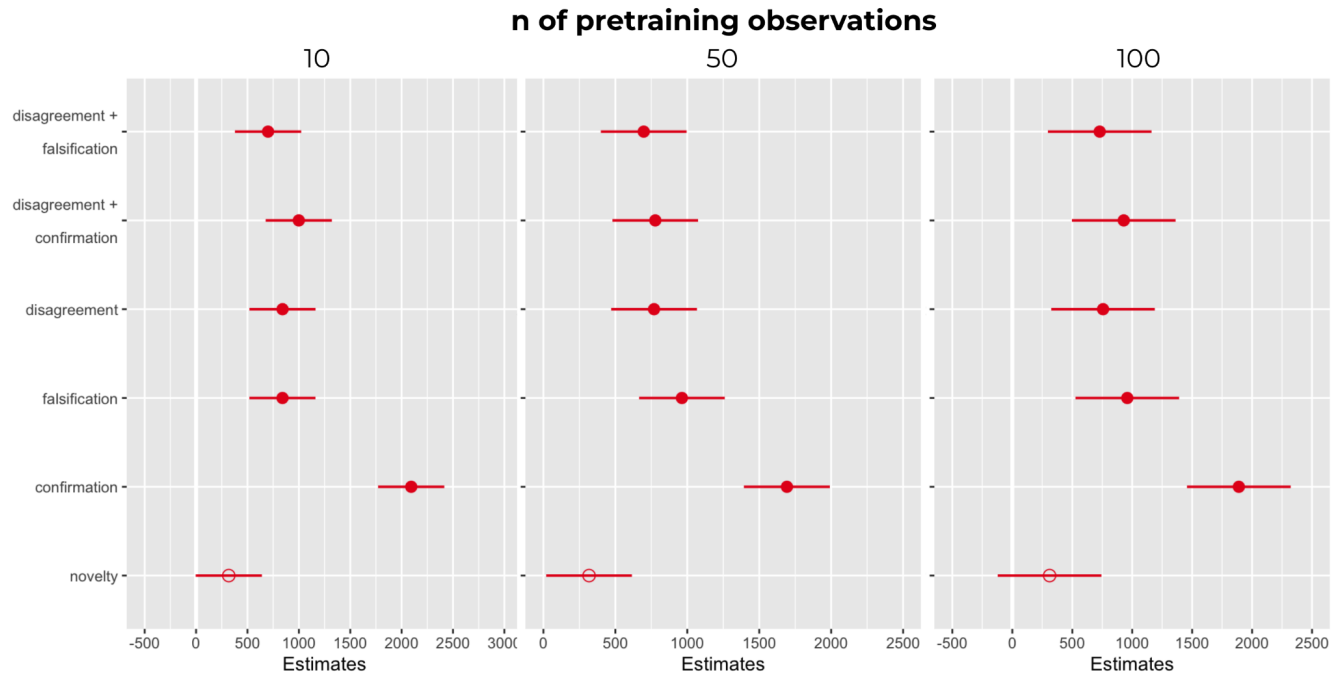

Figure 5. Estimates of experimentation strategies' effects on agents' objective theoretical error (Model 1) for different amounts of pretraining observations per agent. All the coefficients are obtained relative to the baseline random strategy: positive coefficients indicate that the theoretical (reconstruction) error is higher than the baseline (random experimentation), negative coefficients indicate lower than the baseline (random experimentation) error. Notice that the ranking of strategies' effects does not change with the amount of pretraining. Filled circles: $p<0.001$ significance of the effect, transparent circles: $p>0.05$.

3. Discussion

Scientific theories influence science in many ways. They underlie scientific communication, allow scientists to generalize their results to new situations, and guide experimentation within the subfields. Finally, theories often enable scientists to constrain the measurable space by pointing at the dimensions of interest and their relevant variation. The simulations presented here bear on only one of these functions: we test whether it is helpful to alter experimentation based on current theoretical accounts of the phenomena (their successes, failures, or disagreements). Intended to test only one of the functions of theories, the model features some of their other functions: agents develop theories to capture the regularities of the

complex data and communicate the results. Moreover, the measurable space of phenomena that the agents deal with is constrained, potentially reflecting another function of theories that our results are agnostic to.

Our results demonstrate that collecting new data at random is by far the best strategy for learning about the "ground truth". Collecting new observations that are as different from previous observations as possible, is the second best strategy. The falsification-based and crucial experiment-based strategies, often suggested as academic "gold standards", end up generating misleading data for the agents. Finally, the confirmation-based strategy, potentially most commonly used in the academic world, results in the least successful learning even about the simplest ground truth distributions. The relative efficiencies of different experimentation strategies were strikingly invariant to the contexts we varied: the random and novelty-based experimentation were superior to all other tested strategies for ground truths of different complexity, across the learning time, for different group sizes and social learning schemes, resource limitations, and other conditions.

Interestingly, the picture reverses when the agents are evaluated with respect to "subjective" metrics of epistemic success. The confirmation-, crucial experiment-, and falsification-driven scientists generate an illusion of epistemic success, simply by collecting the easiest data to explain. The agents collecting new data at random or in a novelty-seeking way, appear to be the least epistemically successful when assessed against the data that they have access to, because they end up collecting more challenging data that actually reflects the ground truth. Thus, there is a striking inconsistency between the seeming and actual epistemic success: the agents following the least objectively successful experimentation strategies, end up with the highest self-confidence in their own success (this presents many results as in Stewart & Plotkin, 2021 or Rzhetsky et al., 2015 in a new light: these works only study the apparent success, but make conclusions about the objective success).

3.1. The values of randomization

Our main result is quite straightforward: to learn about the ground truth, one should collect as unbiased observations from the ground truth as possible. Note that we already consider random data collection as a gold standard within some aspects of science: for example, we try to randomize samples along the potentially confounding variables within each experiment. For example, when trying to estimate the average human height, we try to randomize the location and age of the participants as much as we can. While randomization is a commonly desirable strategy for the vast majority of statistical estimation problems, we somehow consider it irrelevant to the data collection on a more general level of experimental design.

Random sampling contrasts with the active (biased) sampling strategies that have been demonstrated to enable more efficient learning in a variety of contexts. Namely, in any structured environment, there are ways to speed up theory-building by biasing observation sampling (Settles, 2009; Myung, Cavagnaro & Pitt, 2013). For example, if 16-year-olds have the most close-to-average height, a scientist can estimate average human height much faster if she only samples from this age group. The adaptive sampling, however, requires a great insight into the problem structure that the agents are trying to learn about in the first place. The successful "active learning" strategies facilitate the learning process only in very specific circumstances, such as when the learner starts with an accurate prior knowledge about the problem space, and they have to be used cautiously: only when the learning context corresponds to the one that the strategy has been proven to work on (e.g. Sugiyama, 2005; Sloman et al., under review). In other words, adaptive sampling is *fragile*: if non-random sampling is applied but the original assumptions about the target domain are wrong, or if the strategy is applied imperfectly, the scientist is very likely to be led astray. For example, the average human height inferred from a subsample of 16-year-old people can be any different from the actual average human height whenever the original assumptions for biasing sampling in this way were not entirely correct.

Finally, the active sampling strategies that have been proven successful in specific circumstances are very specific algorithms (e.g. as reviewed in Settles, 2009) that rarely correspond to any of the "theory-motivated" experimentation strategies that scientists strive to, use, or could actually implement.

3.2. Recommendations, limitations and further directions

Due to the generality of the model, the evidence against biasing experimental choice based on theoretical considerations could be interpreted across the scales: as reflecting dynamics within a small subfield or science as a whole. These results question prioritization of theory-driven experimentation among funders, publishers, and editors in many scientific fields. Importantly, there might be reasons to prefer theory-driven experimentation that we have not explored: for example, theory-motivated experiments might be easier to perform, faster, or lead to an easier publication of the results. These reasons prevent us from recommending individual scientists to avoid theory-motivated experiments in each particular case, but urge them to reconsider the commonly assumed virtue of theory-driven experimentation.

It is not clear whether a perfectly random choice of experiments is possible. The closest to ideal random experimentation might correspond to the automatized methods of experimental design that are currently being developed (see Yarkoni et al., 2020). However, even the automatic experimental design software has to operate on the dimensions and variables that are constrained, by theories or any other considerations (see this issue being addressed in Hoffrage & Hertwig, 2006). Potential impossibility of perfectly random experimentation does not undermine the recommendations based on our results. While already being constrained and shaped by theories in many ways, scientists still can choose whether to perform a more or less theoretically motivated experiment in almost each case. For these situations, our work gives an unambiguous answer.

Importantly, the current way of estimating the agents' performance did not take into account the possibility of the agents specializing in particular regions of the ground truth (Weisberg & Muldoon, 2009): all the agents were evaluated against the whole space. Even though this approach is informative as the first brushstroke onto the fundamental question of experimentation choice in science, it eliminates a possibility of a group that ends up learning about the world efficiently through specializing. In fact, some of the strategies end up producing more heterogeneous agents (e.g. confirmation-based), which may prove more efficient if each agent is only evaluated against its area of expertise. We note this as a limitation and a direction for future work, also suggesting that the current results are representative of many realistic scenarios in science. In particular, even within the extremely specialized subfields of scientists' primary expertise (e.g. perception of color in human adults), scientists deal with the phenomena that contain more than one effect or regularity. Therefore, even learning in such very specialized areas that science would involve the strategies that could efficiently deal with the phenomena coming from several distinct structures or families (clusters), and this is exactly the type of process that we investigate in this work, and that shows a failure of all theoretically-motivated experimentation strategies that we tested.

Beyond the individual results reported here, our modeling framework provides a method for evaluating a variety of aspects of scientific practice. We have focused on one particular issue, but the model is sufficiently general to apply to many others including the method of theory construction, how results are communicated, and strategies for dividing labor in a scientific community.

3.3. Conclusion

We present evidence that random experimentation allows agents to develop more successful theories about the ground truth, than a variety of theory-driven experimentation procedures, such as falsification, confirmation, and crucial experimentation. Despite being

objectively superior across all contexts we tested it on, the random experimentation appears to produce inferior theories to scientists themselves, precisely because it prevents agents from inadvertently simplifying their theoretical job by sampling simple, but misleading, observations. We are suggesting that, *when it comes to experimental choice*, scientists should be less influenced by theories than has often been supposed by philosophy of science and by the scientists themselves.

4. Methods

4.1. Model of collective learning

We developed a minimal multi-agent model that captures three essential aspects of scientific activity: collecting the data, building explanations, and learning from others. In the simulations, the agents are trying to develop efficient lower-dimensional representations (theories) of the simulated high-dimensional "ground truth" environments.

4.1.1. Ground truth

On each simulation, we seed a ground-truth environment for the agents to learn about. Each environment is a mixture of multivariate gaussian distributions (N_clusters=[2,10,30]) that span across many dimensions (N_dim=[20;100]; dim_length=200). The ground truth space is used to produce observations for the scientists, which correspond to the points sampled from the overall ground truth distribution. Observations are therefore described by their values along the ground truth's dimensions. Some values along the dimensions are more likely than the others; some dimensions correlate with each other, while the others do not; some observations are clustered around a particular range of values along the dimensions. Therefore, different values along dimensions have differential predictive ability of the values along other dimensions, and some observations are more useful to learn about than others. For example, observing a datapoint from a highly frequent, but narrow, cluster might help generalize to other

observations in the same cluster, but would be uninformative (or even misleading) with respect to the observations generated from other clusters. On the other hand, obtaining a rare and unrepresentative observation might not allow for any generalization at all.

4.1.2. Agents

The agents perform two epistemic actions: 1) collecting the data by controlling a value along one of the dimensions of the ground truth distribution and recording the values along other dimensions of the resulting observation (measurement_capacity=[all the dimensions, half of them], when limited to recording half of the dimensions, the agents choose them randomly; spatial representation of agents' experimentation and theorizing is analogous to Klahr & Dunbar, 1988); 2) constructing compact "explanations" for the collected data. Agents' explanations are lower-dimensional representations of their data, and the agents' goal is to come up with the most efficient and representative lower-dimensional account for the ground truth space. Efficient lower-dimensional representation of the ground-truth essentially allows the agents to predict new observations and the values that they take along the unrecorded dimensions. Here, agents' explanations take the form of neural autoencoders (Kramer, 1992) with one hidden layer (N_hidden_neurons=[3,6]) which are frequently updated on the agents' individual collected data, in the same way for all the experimentation strategy conditions. When the agents update their explanations (see collective strategies for different updating schemes), they undergo 50 stochastic gradient descent updates (batch size=1; optimizer=adam; loss function=mean squared reconstruction error) on their observations serving as both inputs and outputs. The activation function for the agents' autoencoders is chosen to be linear for the encoding layer and ReLU for the decoding layer.

4.1.3. Experimentation strategies

The main focus of this work is the relative epistemic success of different experimentation strategies. We tested the following strategies:

1. **Confirmation**. The agent computes the "theoretical error" (reconstruction error of its neural autoencoder) on each observation in its dataset. Reconstruction error reflects how much information about an observation is captured in an agent's autoencoder (theory). Low reconstruction error on an observation means that this high-dimensional observation can be successfully recovered from agent's lower-dimensional representation of it alone.

   The confirmation-driven agent samples a new observation close to a previous observation chosen with respect to the theoretical error on it: the probability of an observation to be chosen as a "reference" for the new experiment is inversely related to its theoretical error on it (the better the current theory accounts for the observation, the more likely this observation is chosen as a reference for new experiment). When collecting the data in the theory-motivated ways, agents only specify the value along one of the dimensions, which is sampled close to the reference observation's value along that dimension. The dimension on which the sampling will be conditioned is randomly chosen from all dimensions of the reference observation. All other values are then conditionally sampled from the ground truth given the controlled dimension and its value. Thus, the confirmation-driven agents basically follow the gradient of their theoretical success when sampling new observations.
2. **Falsification**. The agent computes the "theoretical error" (reconstruction error of its neural autoencoder) on each observation in its dataset. Then, the agent samples a new observation close to a previous observation chosen with respect to the theoretical error on it: the probability of an observation to be chosen as a "reference" for the new experiment is positively related to its theoretical error on it (the worse the current theory

accounts for the observation, the more likely it is chosen as a reference for new experiment).

3. **Disagreement** (**crucial experiment**). Two agents combine their datasets and generate predictions for each sample in the combined dataset. The observations are ranked according to the euclidean distance of the agents' predictions on them. Then, one of the agents samples a new observation close to the observation from the combined dataset chosen proportionally to its place in the "disagreement" ranking (the more agents disagree on an observation, the more likely it is to become a reference for the next experiment).
4. **Novelty**. The agent randomly generates 500 potential sampling locations, which correspond to the dimension to control and the value along that dimension. The agent computes the distance of each of these potential experiments from all the observations it already collected. Proposals concerning dimensions that have never been observed before (when the agents have a limited measurement capacity) receive a distance of infinity. The next experiment is chosen as the one having the highest minimal distance from already collected samples.
5. **Random**. The agent chooses the conditioning dimension and its value to sample from with uniform probability, independently of its "theory" for the ground truth or previously collected observations.

We also formalized two hybrid strategies, "**confirmation** + **disagreement**" and "**falsification** + **disagreement**", which operate by ranking the observations based on both their disagreement scores and confirmation/falsification scores.

For all strategies except for random, the agents start with fully random sampling up until they collect 10 first observations (unless otherwise specified – see 2.2.). After this initial stage, each agent has a 10% constant random exploration rate.

4.1.4. Social learning strategies

We varied the social learning strategies that a group of agents (N=[5,10]) follows while learning about the world:

1. **Data sharing**. The agents share their observations with each other after they collect each datapoint. Each agent adds another agent's observation to its dataset and updates its explanation on the new dataset each time it gets bigger. Each agent treats its own collected observations and other agents' observations indistinguishably.
2. **Dimension sharing**. Every ten steps, two agents are randomly chosen to exchange how much they value different dimensions of the ground truth. The agents recombine their internal nodes so that one internal node's encoding weights are completely replaced with encoding weights of a randomly chosen internal node from another agent. After recombination of weights, both agents' explanations are retrained on their datasets.
3. **Explanation sharing**. Two agents are chosen every 10 steps to exchange their explanations. Both agents are assigned a pairwise average of their autoencoder weights as their new explanation. After this exchange, the agents' explanations are retrained on their datasets.
4. **Aligned explanation sharing**. Two agents are chosen every 10 steps to exchange explanations. After the exchange, both agents end up with a pairwise average of the weights of their aligned explanations as their new explanation. For this, the agents' internal nodes are first aligned based on the similarity of the weights that the internal nodes have to the input/output nodes. We use the Hungarian algorithm (Kuhn, 1955) to find an alignment of two agents' internal nodes that minimizes the difference between their internal nodes' weights. Then, we take the pairwise average for the aligned nodes' weights and assign it as a new explanation for both agents. The agents' explanations are retrained on their datasets after sharing.

5. **Skeptical explanation sharing**. This social learning strategy is the same as aligned explanation sharing, but the explanation exchange is weighted: the agents' value their own explanations more than the other agents' explanations. For this, we take a weighted (0.8 to own weights, 0.2 to other's weights) average of two agents' aligned weights when they exchange their explanations. The agents are retrained after their weights are reassigned.
6. **Explanation sharing with some data sharing**. The agents share their explanations in an aligned way every 10 experimentation steps, but they also share each observation they collect with one randomly chosen colleague each time they acquire it. The colleague adds the shared observation to its dataset and updates its explanation on its full dataset.
7. **Teaching and learning**. After each experimentation step, one agent is chosen to teach one other agent, imagining new observations based on the data it has seen so far. Formally, the new observation is created as follows. First, the new "raw" observation is made as a linear combination of datapoints that the teacher already collected. The weights in this linear combination are always positive, sum to one and are sampled from a dirichlet distribution. Then, as the last step, this "raw" observation is passed through the teacher's conceptual framework (autoencoder) and reconstructed, thus generating the final observation. The student agent adds the new observation to its dataset and updates its explanation on its full dataset.

Importantly, we vary the collective learning strategies to test the dependence of the experimentation choice strategies' efficiency on these different social contexts. In our simulations, the social interaction strategies are not controlled by the number of learning updates that each agent undergoes or the number of raw datapoints that each agent has in its possession. Therefore, we urge caution in interpreting any differences in performance for the groups with different social learning strategies.

4.1.5. Goal of the epistemic activity and evaluation

Even though the modeling framework is very flexible, in current simulations we explore the results with respect to constructing good theories, or more precisely in our model: efficient lower-dimensional accounts for the underlying reality.

To evaluate scientists' accounts for reality, we compute the reconstruction error that the scientists' theories achieve on their own collected datasets ("subjective") and on 10000 samples from the original "ground truth" distribution ("objective"). Reconstruction error basically reflects the amount of structure in the agent's collected data or the ground truth that the agent's theory captures. In other words, reconstruction error shows how accurately the rich ground truth (objective) or collected data (subjective) can be reproduced given the concise theory alone. We record the average subjective and objective performance of the group, and also the best individual's "objective" and "subjective" performance after every 50 experimentation steps (unless otherwise specified – see 2.2.) until the group collects 300 observations in total.

4.2. Hyperparameters

Even though we tried to vary as many potentially relevant components of the model as possible, many parameters that we did not vary may affect the results. First of all, our simulations end when the group of agents collects 300 observations. Moreover, we had to stick to one arbitrarily chosen updating regime for the agents (50 adam steps, with batch size=1). All these hyperparameters remained constant while we varied the experimentation strategy, and might have affected the generalizability of the reported results.

4.3. Analysis of theoretical heterogeneity

We designed two metrics to assess theoretical heterogeneity of the agents for each simulation:

1. **Heterogeneity of theory-based predictions:** we sample 10000 observations from the ground truth. We run each of these observations through the agents' final theories to obtain their predictions. Then, we compute the average pairwise euclidean distance of agents' predictions for each observation and average them out to obtain the group's heterogeneity score.
2. **Heterogeneity of the theory representations:** we compute the mean euclidean distance between the weights of all pairs of agents in a group after aligning the internal nodes of each pair with the Hungarian algorithm (Kuhn, 1955).

4.4. Analysis of sampling behavior

We designed a number of metrics to shed light onto the sampling behavior of the agents for each simulation:

1. **Individual sampling variability:** for each agent, we compute pairwise euclidean distance between all its observations. Then, we compute an average of the sampling variability for each group as a measure of individual variability of samples for each simulation.
2. **Individual sampling variability over time:** we look at how the variability of samples changes over time. We compute sequential pairwise euclidean distance of the samples in each agent's dataset. Then, we average the sequential distances between the agents in the group, and compute a spearman correlation of the average sequential distances with their order (time).
3. **Between-agent sampling variability:** we compute an average pairwise euclidean distance between the samples of all agents in a group.
4. **Between-agent sampling variability over time:** we compute an average pairwise euclidean distance between agents' observations sampled at the same order (time). Then, we compute the spearman correlation of the sequential between-agent distance of samples with order of the observation.

5. **Representativeness of samples:** we compute per-group average log likelihood of sampled observations in the ground truth.
6. **Representativeness of samples over time:** we compute the spearman correlation of the average log likelihood of the observations collected by the agents in a group at each step with time.

Acknowledgements

The authors thank Sabina Sloman, James Michelson, Robert Goldstone, Mahi Luthra, Joshua Nunley, Iain Cunningham, members of PCL, Computational Neuroethology, and Building a Mind labs at Indiana University, participants of the Pittsburgh Formal Epistemology Workshop for productive discussions that lead to the improvement of this work. M.D. was supported by IU Cognitive Science program, CMU Center for Formal Epistemology, and the NSF-NRT grant 1735095 "Interdisciplinary Training in Complex Networks and Systems". This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute: https://kb.iu.edu/d/anwt\#carbonate.

References

1. Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612.
2. Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193.
3. Fleck, L. (1979). *Genesis and Development of a Scientific Fact*.
4. Hoffrage, U., & Hertwig, R. (2006). Which world should be represented in representative design. *Information Sampling and Adaptive Cognition*, 381–408.
5. Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389.
6. Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press on Demand.
7. Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1–48.
8. Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211.

9. Kramer, M. A. (1992). Autoassociative neural networks. *Computers & Chemical Engineering*, *16*(4), 313–328. https://doi.org/10.1016/0098-1354(92)80051-A

10. Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1–2), 83–97.

11. Kuhn, T. S. (1970). *The structure of scientific revolutions* (Vol. 111). Chicago University of Chicago Press.

12. Lakatos, I. (1974). The role of crucial experiments in science. *Studies in History and Philosophy of Science Part A*, *4*(4), 309–325.

13. Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

14. Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*(3–4), 53–67.

15. Palmer, D. S., Howrigan, D. P., Chapman, S. B., Adolfsson, R., Bass, N., Blackwood, D., Boks, M. P. M., Chen, C.-Y., Churchhouse, C., Corvin, A. P., Craddock, N., Curtis, D., Di Florio, A., Dickerson, F., Freimer, N. B., Goes, F. S., Jia, X., Jones, I., Jones, L., … Neale, B. M. (2022). Exome sequencing in bipolar disorder identifies AKAP11 as a risk gene shared with schizophrenia. *Nature Genetics*, *54*(5), 541–547. https://doi.org/10.1038/s41588-022-01034-x

16. Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. routledge.

17. Rzhetsky, A., Foster, J. G., Foster, I. T., & Evans, J. A. (2015). Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, *112*(47), 14569–14574.

18. Settles, B. (2009). *Active learning literature survey*.

19. Shi, C., & Tang, B. (2021). Model-Robust Subdata Selection for Big Data. *Journal of Statistical Theory and Practice*, *15*(4), 1–17.

20. Sloman, S. J., Oppenheimer, D. M., Broomell, S. B., & Shalizi, C. R. (2022). *Characterizing the robustness of Bayesian adaptive experimental designs to active learning bias* (arXiv:2205.13698). arXiv. https://doi.org/10.48550/arXiv.2205.13698

21. Stewart, A. J., & Plotkin, J. B. (2021). The natural selection of good science. *Nature Human Behaviour*, *5*(11), 1510–1518.

22. Sugiyama, M. (2005). Active learning for misspecified models. *Advances in Neural Information Processing Systems*, *18*.

23. van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

24. Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, *76*(2), 225–252.

25. Yarkoni, T., Eckles, D., Heathers, J. A. J., Levenstein, M. C., Smaldino, P. E., & Lane, J. (2021). Enhancing and Accelerating Social Science Via Automation: Challenges and Opportunities. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.df2262f5

26. Young, M. E., Cole, J. J., & Sutherland, S. C. (2012). Rich stimulus sampling for between-subjects designs improves model selection. *Behavior Research Methods*, *44*(1), 176–188.

27. Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis*, *72*(1), 17–35.