# Towards Strong AI: Transformational Beliefs and Scientific Creativity

Samuel J. Eschker and Chuanhai Liu

Department of Statistics, Purdue University

April 16, 2025

*The zenith of human intelligence is often portrayed as the ability to create, and to create radically new and/or surprising things.*

— Geraint A. Wiggins (2006)

**Abstract**

Strong artificial intelligence (AI) is envisioned to possess general cognitive abilities and scientific creativity comparable to human intelligence, encompassing both knowledge acquisition and problem-solving. While remarkable progress has been made in weak AI, the realization of strong AI remains a topic of intense debate and critical examination. In this paper, we explore pivotal innovations in the history of astronomy and physics, focusing on the discovery of Neptune and the concept of scientific revolutions as perceived by philosophers of science. Building on these insights, we introduce a simple theoretical and statistical framework of weak beliefs, termed the Transformational Belief (TB) framework, designed as a foundation for modeling scientific creativity. Through selected illustrative examples in statistical science, we demonstrate the TB framework's potential as a promising foundation for understanding, analyzing, and even fostering creativity — paving the way toward the development of strong AI. We conclude with reflections on future research directions and potential advancements.

*Key Words*: Computational creativity; Chain-of-thought; Chain-of-verification; Inferential models; Generalized logic of science;

# 1 Introduction

This paper is the first in a series of three by the authors, exploring statistical reflections on strong AI. Strong AI, as envisioned here, would possess general cognitive abilities and scientific creativity akin to human intelligence, enabling it to approach knowledge acquisition and problem-solving. The other two papers in this series focus on *individual cognition* and *scientific reasoning*, respectively. This explains the title of the present article. Here, we consider scientific creativity from a statistical perspective. For a concrete discussion at the foundational level, we shall focus on scientific inquiry, where science is simply meant to gain knowledge from experience or experiments (see, *e.g.*, Newton, 1718; Martin and Liu, 2015a). As it is often the case that science inquiry is dynamic (Popper, 2005; Kuhn, 1970), a simple but proper statistical setting can be written as

$$(\Omega_\tau, D_\tau, M_\tau, \Theta_\tau) \tag{1}$$

where $\tau$ indices the dynamic state such as sample size and time, $\Omega_\tau$ the world or environment of interest from which the observed data $D_\tau$ were collected, and $M_\tau$ the model with the space $\Theta_\tau$ of unknown parameters. Perhaps, it can be argued that the success of deep learning (DL) is largely due to its flexibility for a $\tau$-dependent or dynamic approach to modeling data structures. A recent relevant discussion, with the focus on parameter estimation, is provided in Jiang and Liu (2024).

Although somewhat special with respect to the complexity of the real world of scientific inquiry, the statistical setting (1) is deemed adequate to interpret the current logic foundations of weak AI. Here, this paper considers the creativity aspect of strong AI. While it has been playing a fundamental role in scientific inquiry from ancient times, creativity has appeared to be such an elusive concept that it is hitherto difficult to have a well-accepted definition. For example, in their review paper on computational creativity, Carnovalini and Rodà (2020) noticed that researchers analyzed over 200 of definitions of creativity in literature. In the broader context of scientific discovery, philosophers of science have had unsettled debates for centuries on the closely related concept of *scientific discovery*, which can be viewed as the processes and products of scientific creativity. For instance, Schickore (2022) wrote:

> *Philosophical discussions of scientific discovery have been intricate and complex because the term "discovery" has been used in many different ways, both to refer to the outcome and to the procedure of inquiry. In the narrowest sense, the term "discovery" refers to the purported "eureka moment" of having a new insight.*

The term of "eureka moment" or happy moment in the quote marks its importance in the ultimate definition of creativity consistent with our common sense.

As difficult as it may be, a quantitatively meaningful definition of creativity might not be possible in general. However, we strive for a narrow but precise definition of scientific creativity, particularly in the context of scientific discovery where the existing solutions have been found questionable in terms of either validity or efficiency. Our definition of scientific creativity is formulated into a somewhat simplest possible statistical framework of creativity based on both a study of a selected list of great innovations in the history of science in Section 2 and our understanding of the perspective of philosophers of science. This framework is formulated within the context of dynamics for scientific discovery and can be summarized briefly as follows. Firstly, we expand the special setting (1) into a general setting of scientific discovery at the dynamic state $\tau$:

$$(\Omega_\tau, D_\tau, M_\tau, \Theta_\tau) \stackrel{\text{Create}}{\longrightarrow} (\Omega_{\tau'}, D_{\tau'}, M_{\tau'}, \Theta_{\tau'}) \tag{2}$$

where $\Omega_{\tau'}$ is a new world or population made of the original $\Omega_\tau$ and potentially additional some auxiliary world $\Omega_{\text{mis}}$, $D_{\tau'}$ is the new data, and $M_{\tau'}$ is the new model with the space of $\Theta_{\tau'}$ to be considered to address the observed discrepancy in the scientific inquiry. Secondly, creative ideas are entertained iteratively or in parallel, according to some iteration among the following three steps: *Creation* constructing $(\Omega_{\tau'}, D_{\tau'}, M_{\tau'}, \Theta_{\tau'})$ in (2); *Exploration* entering a necessary stage that Kuhn (1970) calls *normal research*; and *Evaluation* comparing the *predicted* against the *observed* as a way of evaluation, confirmation, or verification. When successful, creative solutions are found in the sense that can be considered as the purported "eureka moment" of having a new insight. Within this context, our scientific creativity is defined as the transforming procedure of Creation subject to the verification by the Evaluation step. In other words, we consider it creative, as it creates a new world to solve the unsolved problem. We call the above statistical approach the transformational belief (TB) framework of scientific creativity, with resulting new TBs obtained with respect to the new world indexed by $\tau'$.

The rest of the paper is carried out itself in an inductive fashion and is arranged as follows. Section 2 develops intuitions for a definition of scientific creativity by studying both a selected list of great innovations in the history of astronomy and physics and the perspectives of philosophers of science on scientific discovery. Section 3 presents the TB framework of scientific creativity in a dynamic data-driven environment for problem-solving. Section 4 illustrates TB with a relatively simple artificial example, a many-normal-means problem, showcasing the potential of TB for strong AI, while Section 5 demonstrates it with the discoveries in the attempts to develop inductive inference, also known as the logic of science (see, *e.g.*, Jaynes, 2003). For the latter, a simple experimental computational evaluation is also considered using ChatGPT (OpenAI, 2024), a large language model (LLM), with a manually dynamic chain of thought and verification. Section 6 concludes with a few remarks for future research.

# 2 Historical Discoveries in Natural Science

In this section, we develop intuitions for our definition of creativity by studying some examples of great discoveries in astronomy and physics. Our exposition aims to provide sufficient details on the discovery of Neptune in Section 2.1 and to summarize our investigations on a list of other great discoveries in the history of astronomy and physics in Section 2.2. Section 2.2 also contains a brief review of the perspectives of philosophers of science on scientific discoveries and revolutions, which more or less agree with our observations.

## 2.1 The groundbreaking discovery of Neptune

A potential application of exploratory and transformational beliefs pertains to a scientist's reasoning process, through which they *accept or reject* their theories based on observed evidence. This uses probability theory and has been extensively debated by philosophers for over a century (see, *e.g.*, Jaynes, 2003, p. 133, §5.5). For a specific example, we consider the discovery of Neptune and in a chronological order, our exposition begins with the discovery of Uranus.

### 2.1.1 The discovery of Uranus

Aside from Earth, five planets — Mercury, Venus, Mars, Jupiter, and Saturn — are easily visible to the naked eye and have been known since ancient times. Uranus was the first planet discovered with the aid of a telescope by William Herschel on March 13, 1781, while conducting a systematic sweep of the contents of the night sky. Initially, he believed that he had found a comet because the object appeared to move relative to the stars. However, further observations by Herschel and other astronomers revealed that the object had a nearly circular orbit around the Sun at a distance about twice that of Saturn (see the Titius-Bode law (Gregory, 1715)), suggesting that it was a planet rather than a comet, which would have had a highly elliptical orbit.

Herschel's finding extended the boundaries of the solar system and marked a significant advancement in astronomical research. For example, earlier star catalogs revealed that Uranus had been observed 20 times before its identification as a planet in 1781, dating back as early as 1690, but it was mistakenly identified as a star. Even more intriguingly, in 1821, French astronomer Alexis Bouvard compiled all available observations, spanning a period during which Uranus had traversed about one-third (32/84) of its orbit, and encountered a significant issue (Bouvard, 1821). Despite accounting for the gravitational influences of the giant planets Jupiter and Saturn, he was *unable to reconcile the observed data with predictions* based on the Newtonian theory, laws of motion and gravitation.

### 2.1.2 The discovery of Neptune

The irregularities in the orbit of Uranus led Bouvard to hypothesize some perturbing body. The irregularities, both in the planet's ecliptic longitude and in its radius vector, could have been explained by several hypotheses, including:

$H1$. the effect of the Sun's gravity at such a great distance might differ from Newton's description, *i.e.*, the Newtonian theory was demolished;

$H2$. the discrepancies might simply be observational error; or

$H3$. perhaps Uranus was being pulled, or perturbed, by an undiscovered planet or multiple planets that are farther away from Uranus (Grosser, 1964).

Like Bouvard, French astronomer Urbain Jean Joseph Leverrier (1811-1877) and English scholar John Couch Adams (1819-1892) from St John's College, Cambridge, likely regarded $H3$ as more plausible than $H1$, $H2$, and potentially many other hypotheses. This is evident from the immense computational efforts they invested in further pursuing their mathematical search of the hypothetical disturbing planet. Working within the framework of H3, Adams and Verrier independently reached a predicted position for the hypothesized perturber on the celestial sphere, using essentially the same perturbation theory techniques and the Titius-Bode law. On 24 September 1846, Johann Gottfried Galle and Heinrich Louis d'Arrest at the Berlin Observatory spotted a new planet that is very close to the predicted position by Verrier, who named this planet of the Solar system *Neptune*. This is a truly remarkable achievement for the epoch, as Airy (1846) wrote (p.121):

> *In the whole history of astronomy, I had almost said in the whole history of science, there is nothing comparable to this. The history of the discoveries of new planets in the latter part of the last century, and in the present century, offers nothing analogous to it.*

In modern terms, the problem tackled by Adams and Verrier is an inverse problem. In a recent revisit to the problem, Rodríguez-Moris and Docobo (2024) recomputed the perturbations induced in the orbit of Uranus by Neptune, using the data from Solar System Dynamics at `https://ssd.jpl.nasa.gov`.

#### 2.1.3 Statistical evidence for patterns of scientific discovery

It is remarkably worth noting that the process in the discovery of Uranus begins with an abnormal phenomenon that contradicts the prediction principle: *the observed is inconsistent with the expected or predicted*, providing clear evidence that demands further investigation. The same phenomenon occurred again with the discovery of Neptune. Perhaps more importantly, our observations suggest the possibility of statistical modeling of scientific discovery.

The above observations no doubt shed light on a meaningful definition of scientific creativity. Typically, creative innovations come next when investigators conduct a new scientific investigation to resolve the discovered anomaly. In the case of the discovery of Uranus, astronomers weakened their previous beliefs, and thus established new beliefs by remodeling the observed data. In the case of the discovery of Neptune, the beliefs were transformed, followed by remodeling the orbit data Uranus with missing values, the hypothetical planets.

### 2.2 More examples of great innovations in astronomy and physics

Far more historical examples are available than we have had space to exploit here. The observations of our studies of great discoveries in celestial mechanics and physics from a statistical perspective of their innovations are briefly summarized in Supplementary A. Great discoveries all typically start with experiments and observations, build mathematical theories or statistical models, verify the new theories with experiments and new observations, and iterate such a process towards further verification, improvement, and discoveries.

Extensive existing research on the science of creativity and discovery, primarily conducted by philosophers of science, has also been undertaken (see, *e.g.*, Kuhn, 1970; Aleinikov, 2013; Schickore, 2022, and references therein). Their perceived general structure or pattern of such activities, particularly in natural science, exhibits similarities to our observations in this section. From a statistical perspective, we formulate in the next section a framework of scientific discovery in environments characterized by such a common process.

## 3 Transformational Beliefs: a General Framework

### 3.1 The prediction principle

Although the focus here is on scientific creativity, our discussion cannot be independent of the scope of scientific discovery for problem-solving in scientific inquiry. It can arguably be said that one of the first principles of science is the principle that the observed data and the predicted data must be consistent with each other. The general idea can be traced to Isaac Newton (1704, Newton (1718) and Schickore (2022)), as seen in his method of analysis:

*"As in Mathematicks, so in Natural Philosophy, the Investigation of difficult Things by the Method of Analysis, ought ever to precede the Method of Composition. This Analysis consists in making Experiments and Observations, and in drawing general Conclusions from them by Induction, and admitting of no Objections against the Conclusions, but such as are taken from Experiments, or other certain Truths ... By this way of Analysis we may proceed from Compounds to Ingredients, and from Motions to the Forces producing them; and in general, from Effects to their Causes, and from particular Causes to more general ones, till the Argument end in the most general. This is the Method of Analysis"*

We refer to the fundamental underlying principle as the *prediction principle* . This principle was made clear in William Whewell's view, as Snyder (2023) summarized:

*On Whewell's view, once a theory is invented by discoverers' induction, it must pass a variety of tests before it can be considered confirmed as an empirical truth. These tests are prediction, consilience, and coherence (see Whewell, 1858, p. 83-96). These are characterized by Whewell as, first, that "our hypotheses ought to fortel [sic] phenomena which have not yet been observed" (Whewell, 1858, p. 86); second, that they should "explain and determine cases of a kind different from those which were contemplated in the formation" of those hypotheses (Whewell, 1858, p. 88); third that hypotheses must "become more coherent" over time (Whewell, 1858, p. 91).*

Notably, using modern inductive inference in terms of a sound logic of science (see Section 5), we can arguably perform Whewell's tests of significance against *consilience* and *coherence* based on the more fundamental concept of *prediction*.

With the prediction principle, here in this section we consider a narrow but concrete definition of scientific creativity that is summarized the three-step TB framework introduced in Section 1.

### 3.2 A philosophical perspective

The three-step idea here is similar to or, to some extent, can even be viewed as a modern logic of science-based renovation of that of scientific discovery of three steps of Whewell (1840); see the discussion in Schickore (2022): the happy thought or 'eureka moment', the articulation and development of that thought, and the testing and evaluation of it. While there are diverse opinions of philosophers of science on possible logic of discovery, we focus on a pragmatic approach by considering the context of problem-solving as the setting. Philosophically, this is dfferent from W. Whewell, as, for example, Schickore (2022) wrote: '*According to Whewell, the initial step in every discovery is what he called "come happy thought, of which we cannot trace the origin; some fortunate cast of intellect, rising above all rules. No maxims can be given which inevitably lead to discovery" (Whewell 1996 [1840]:186).* ' To the extent of simulating happy thoughts, especially in the current prevailing era of generative-AI, this seems to be consistent with the common sense understanding of creativity, *i.e.*, generating a vast array of unconventional hypotheses. This is also consistent with the idea in the literature that the special logic of discovery is the logic of abductive or "retroductive" inferences (Hanson, 1958, 2014; Schickore, 2022), which we refer to in Section 3.3 as pragmatic reasoning of "reverse-engineering".

However, it is critical to have deep reflections when it comes to thinking about strong AI. In reality, it is important to observe that having the happy thought in discovery is unlikely to be a wild and capricious guess (see Schickore, 2022, and references therein for similar opinions). As in the studies shown in Section 2, scientific discoveries often involve noticeable inconsistencies between experience and theory. Actions are then necessary to change beliefs that have been obtained from a logic of science based on the current propositions, the data, and the resulting model. This explains the rationale for our proposed approach, called the framework of *exploratory and transformational beliefs* or, simply, the *transformational belief* (TB) framework.

## 3.3 Abduction, reverse-engineering, and the TB framework

Our discussion of scientific creativity for strong AI is narrow and thus could be said to be an inquiry of statistical discovery. Recall that in general, problem-solving is an iterative process until certain conditions implementing the prediction principle are satisfied. Thus, in the context of statistical modeling of scientific creativity, the basic idea can be summarized as: at a simplest level, all your need is to consider *reverse-engineering, re-sampling, and re-modeling;* following Albert Einstein:

> *It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.*

This is often quoted as 'Everything should be made as simple as possible, but not simpler', 'On the Method of Theoretical Physics', lecture delivered at Oxford, 10 June 1933; See Ratcliffe (2014).

Technically, our jump point is therefore the setting for such a dynamic environment introduced as (1), where for each iteration, we have an unsolved problem raised by verification of the observed and predicted. With this setting, we discuss the three steps of each iteration in detail below.

*The creation step.* Given the unsolved problem, creative approaches relay on the appropriate experience or data to be created to solve the problem. This can include re-sampling and retrospective reconstruction (or reverse-engineering) of data available for the previous analysis. It may also include approaches of retrospective reconstruction or reverse-engineering to select or remove observations. The goal is that the selected or remaining data with the new data, denoted by $D_{\tau'}$, would be used via adequate remodeling to produce valid inference. Consequently, we have the population underlying $D_{\tau'}$, which is the resulting population or the explicitly targeted population that represents the environmental context of the problem to be solved. This leads to a dynamical modification of (1) to (2) and, thereby, the *transformed beliefs*.

*The exploration step.* This step, also known as the articulation and development step, aims to advance research based on the new theory or model by formulating and disseminating mathematical or statistical hypotheses about quantities of interest, grounded in scientific logic. It facilitates ongoing research, leading to refined beliefs, including more broadly exploratory ones. When necessary, these probabilistic findings can lead to the subsequent evaluation step with numerical inputs.

*The evaluation step.* This step, also known as the verification step, applies the prediction principle to compare *predicted* against *observed* as a way of evaluation, confirmation, or verification. At a technical level, this step involves formulating assertions or hypotheses on the validity of the current model or system of beliefs and generating uncertainty assessment based on the current model or system of beliefs. When strong evidence arises against the adequacy of the current theory or model, the next Creation step is called for.

Given the creation and evaluation steps, the exploration step is seemingly redundant but can be viewed as to correspond to what Kuhn (Kuhn, 1970) calls ‘normal’ research. While philosophically similar to Whewell’s structure of scientific discovery and drawing upon Kuhn’s theory of scientific revolutions (Kuhn, 1970), TB is formulated in a principled manner with a sound logic of science to understand and analyze scientific discovery.

Consequently, TB is quantifiable, as demonstrated in Section 4, showcasing the potential of TB for strong AI. It can also be used to evaluate scientific methods and theories, even including inferential theory itself, as shown in Section 5. These examples are intentionally chosen for simplicity and conceptual clarity for general readers.

# 4 A Simple Illustration: Many Normal Means

Scientific discovery involves formulating and, when necessary, refining or rejecting hypotheses about the nature of the world. As we acquire evidence, we create models that we believe capture the structural mechanisms giving rise to our data. Over time, we may refine our parameter estimates for these models, and slowly the models will both fit our existing data better, and have higher predictive capacity. However, we will occasionally encounter situations where new data appear to entirely discredit an existing model. Falling victim to the sunk cost fallacy, these new data may initially be rejected on the basis of their nonconformity, and scientific progress may stall due to a hesitation to explore alternative theories when so much work has already been devoted to determining implications of the existing model. Nevertheless, statistical tools are available for formulating and testing whether an old model should be rejected in light of new data. This is particularly pertinent today. In the modern era of deep learning, foundation models are trained on massive datasets and then fine-tuned for specific use cases, but it is not often clear when a model should be fine-tuned as opposed to outright retrained. A special case, which we explore for the purpose of illustration, is the problem of many normal means, where we ask whether a normal mixture model with a given number of components and previously accepted set of means can reasonably concord with a new observation. This situation arises in repeated experiments, where the question becomes whether we are observing a more diverse population that previously thought. For example, if you were to take note of birds spotted in a prairie and you had four categories of known birds, you may at first mistakenly classify a new, but similar, bird into one of the existing categories. However, as measurement accuracy or sample size increases, it may become clear that the previous four categories are insufficient to describe the population.

More formally, the many-normal-means problem is about making inference on the unknown means $\mu_1, ..., \mu_n$ from the sample $Y_1, ..., Y_n$ with the model

$$\mu_i \sim \mathbb{P}_\theta \quad \text{and} \quad Y_i|\mu_i \sim N(\mu_i, 1) \qquad (Y_i \in \mathbb{Y},\ \theta \in \Theta^{(n)} \subseteq \mathbb{R}^p)$$

for $i = 1, ..., n$ (see, *e.g.* Efron, 2016; Jiang and Liu, 2024). The inferential goal is to estimate $\mu_i$, to model $Y_i$, to predict a new observation, or to produce uncertainty quantification on assertions of interest. As a further simplification of this problem for our illustration purpose, suppose that $\mathbb{P}_\theta$ is a normal mixture model with an unknown number of components, $K$, each with the same known, small variance, $\sigma^2$. To be clear, models with different $K$ values here are meant to be different models in our illustration. Therefore, we have

$$\mu_i \sim \sum_{k=1}^{K} \pi_k N(\phi_k, \sigma^2), \qquad \sum_{k=1}^{K} \pi_k = 1. \tag{3}$$

Using our samples, $Y_i$, we aim to perform inference on $K$, $\pi_1, \ldots, \pi_K$, and $\phi_1, \ldots, \phi_K$ in an iterative sense. Let $g_{n,K}(Y_1, \ldots, Y_n) : \mathbb{R}^n \to \mathbb{R}^{2K}$ be an estimator for $\theta_K \equiv (\pi_1, \ldots, \pi_K, \phi_1, \ldots, \phi_K) \in \mathbb{R}^{2K}$, and let $h_n(Y_1, \ldots, Y_n) : \mathbb{R}^n \to \mathbb{Z}^+$ be an estimator for $K$. We refer to $h_n$ as the transformative level and $g_{n,K}$ as the exploratory level of the model. An iterative procedure would utilize the transformative estimator $h_{n-1}(Y_{-k}) = K_{n-1}$ and its corresponding exploratory estimate $g_{n-1,K_{n-1}}$ along with the new sample, $Y_k$, to *evaluate* hypotheses articulated in the exploration step.

Beyond this one initial sample, the process of gathering new observations creates a time series which requires continual evaluation and, sometimes, creation or exploration. In the case above, we had a single new observation, and we explored our landscape of conclusions under that new observation. Now, suppose that we have a set of observations

$$\{Y_t :\ Y_t \in \mathbb{R};\ t \in \mathbb{Z}^+\}$$

where $Y_{1:n} = Y_1, \ldots, Y_n$ are initial observations and $Y_{n+1}, Y_{n+2}, \ldots$ are future observations. Beginning at time $n$, we must create (creation) an initial model and estimate (articulate) the parameters for that model. Then, at each point $t \in [n+1, \infty)$, we perform an evaluation step to see if the current model reasonably explains the new data. If that evaluation step leads us to reject the model, then we must continue with a new creation and exploration procedure. Next, we outline the details of each of these steps, continuing with the setup in (3).

It should be noted that for our exposition, familiar simple statistical procedures are used. Alternative methods such as those based on inferential models (see Section 5) can be considered, especially in future research in developing TB creativity systems.

## 4.1 Creation

In the creation step, we calculate an initial estimate of $K_n$. This estimate does not need to be optimal, because we can test against other hypotheses in the exploration step, but it should be reasonable in the sense that the model fits the data. One way to estimate the number of components is using penalized maximum likelihood, such as BIC (Schwarz, 1978) where we optimize

$$\text{BIC} = -2\ell(\theta|Y_1, \ldots, Y_n) + K\log(n) = C(\theta) - 2\sum_{i=1}^{n} \log \sum_{j=1}^{K} \pi_j \exp\left\{-\frac{(Y_i - \pi_j)^2}{2}\right\} + K\log(n)$$

by first minimizing $-2\ell(\theta|Y_1, \ldots, Y_n)$ for each $K \in 1, \ldots, n$ separately and using the $K$ which produces the lowest BIC. By generating a sequence $Y_1, Y_2, \ldots,$ we can see how the initial estimate of $K$ changes based off how many samples we have. A demonstration of this is shown on the left hand side of Figure 1. We note that while most of the time, a larger sample provides a more accurate estimate for the true $K$, sometimes this is not

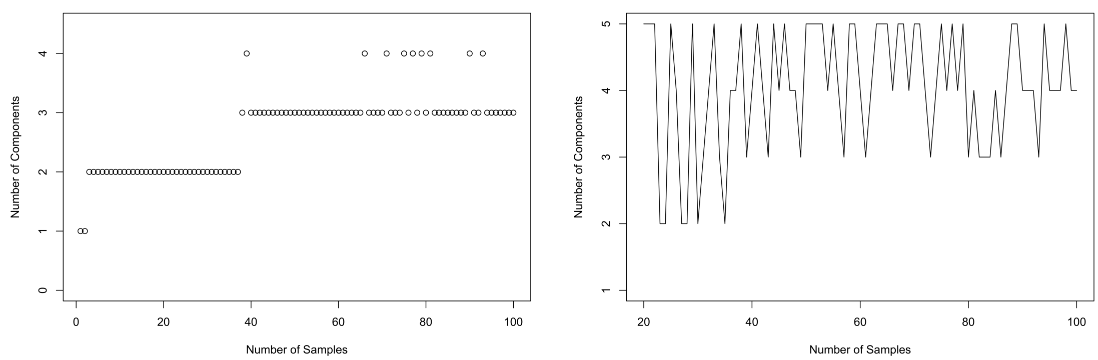


Figure 1: Initial estimates of the number of components by sample size using BIC (left) or 80-20 cross validation of the negative log likelihood (right) when the true model has $K = 3$, $\theta = (-2 \quad 2 \quad 5)^\top$ and $\pi = (0.3 \quad 0.5 \quad 0.2)^\top$.

the case, even for fairly large sample sizes. Now, if instead of using BIC, we used cross validation with 50 replications of an $80 - 20$ split of training and testing, and then chose the number of components with the lowest mean negative log likelihood for each sample size, the behavior is even more oscillatory between various values of $K$, as seen on the right hand side of Figure 1. Both of these methods, however, would provide decent initial estimates for $K$ that are useful.

To estimate $\pi_1, \ldots, \pi_K$ and $\phi_1, \ldots, \phi_K$, two common options are maximum likelihood using the Expectation-Maximization (EM) algorithm or sampling from a Bayesian posterior with a prior distribution

$$\pi \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \qquad \alpha_1 = \alpha_2 \cdots = \alpha_K = 1.$$

For our numerical experiments, we implemented a standard Gibbs sampler and the EM algorithm (see Supplementary B).

## 4.2 Exploration and Evaluation

The evaluation step in scientific creativity is comprised of hypothesis testing and falsification. Given existing or new data, we check whether the model created during the creation step reasonable explains the data in the context of the alternatives articulated above. In this case, we articulate the hypotheses $H_0 : K_{n-1} = K_n$ versus $H_a : K_{n-1} \neq K_n$. If we fail to reject $H_0$, then a new corresponding exploratory estimator $g_{n,K_{n-1}}(Y_1, \ldots, Y_n)$ can be calculated. Alternatively, if we reject $H_0$, we estimate $K_n = h_k(Y_1, \ldots, Y_n)$ and $g_{n,K_n}$, and we say that $Y_k$ catalyzed a transformative discovery.

To take the simplest case, consider that we have a sample $Y_1, \ldots, Y_{n-1}$, and that $h_{n-1}(Y_{-n}) = K_{n-1} = 1$. This implies that $\pi_1 = 1$, and a reasonable estimator for $\phi_1$ is the sample mean, $\bar{Y}_{-n}$. Now, we observe $Y_n$ and test whether $H_0 : K_n = K_{n-1} = 1$ or $H_a : K_n = 2$. Let $0 < \alpha < 1$ be our confidence level, and we reject $H_0$ if $1 - \Phi\left(\left|\frac{\bar{Y}_{n-1} - Y_n}{\sqrt{\sigma^2+1}}\right|\right) < \frac{\alpha}{n}$, where $\Phi(.)$ denotes the cumulative distribution (*c.d.f.*) function of the standard normal distribution. Here we use the Bonferroni adjustment for multiple testing, because each of the $n$ samples is assumed to be *i.i.d.* under the model for $Y_{-n}$. If we fail to reject $H_0$, then we can update our mean estimate as $\phi_1 = \frac{(n-1)\bar{Y}_{n-1} + Y_n}{n}$. However, if we reject $H_0$, then we need to estimate all of $\pi_1, \phi_1$, and $\phi_2$. Note that because $\pi_2 = 1 - \pi_1$, we need not estimate $\pi_2$ directly. Therefore, we are fitting the model

$$f(x; \phi_1, \phi_2, \pi_1) = \frac{1}{\sqrt{2\pi}} \left[ \pi_1 \exp\left( -\frac{(x-\phi_1)^2}{2} \right) + (1-\pi_1) \exp\left( -\frac{(x-\phi_2)^2}{2} \right) \right],$$

which can be done by maximum likelihood using the EM algorithm as described in Supplementary B.

Now, we assume that $n = 10$ and that we previously concluded that our data follow a standard normal centered at their sample mean. That is, $K_{10} = 1$, $\pi_1 = 1$, and $\phi_1 = \bar{Y}$. We analyze what occurs when we observe a new value, $Y_{11} \in (-5, 5)$, and re-estimate the model at both the exploratory and transformative levels. The right-most plot of Figure 2

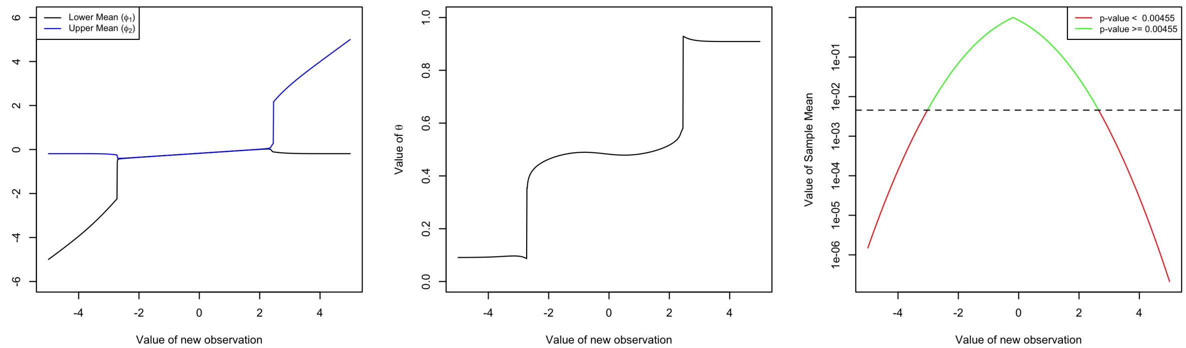


Figure 2: An illustration of the effect of one new observation on model specification for an initial sample with $K_n = 1$, $X_i \sim N(0, 1)$, and $i \in 1, \ldots, n = 10$.

shows the exploratory (green) and transformative (red) regions of observations for $Y_{11}$. If $X_{11}$ falls within the exploratory region, then the model is refined, rather than rejected. However, if the model is rejected (red), then we have additional parameters to estimate. The left two plots show the behavior of the MLE under the new observation given that we reject the current model. These plots showcase a region of stability of the MLE with both estimated means around 0 when the new observation falls close within the range of previously observed values. The shift is reflected in the middle plot as well, which indicates that the model essentially merges from two components back to one when the new observation is indistinguishable in distribution from existing values. In scientific inquiry, analyses such as these are important, because they provide a test-bed for falsifiability. We are asking what data would lead us to reject our current model, and under what reformulating our model would entail given various new observation values.

## 5 The 260-Year Quest for a Unified Logic of Science

There have been too many creative works in statistics that would require a book-long space to discussion (see, *e.g.*, Stigler, 2002; Bickel and Doksum, 2015; Gelman and Vehtari, 2021). In this section, we take a look at the quest for a unified logic of science, the most important unsolved problem in statistics (Efron, 2012), from an exploratory and transformational beliefs perspective.

In Gigerenzer and Murray (2015), the authors argued that in the science of mind, theories are particularly likely to come from tools, and they are specially concerned with the emergence of the metaphor of the mind as an intuitive statistician. Here, tools, physical for changing environment or logical for reasoning, are certainly products of creativity. Their observations are particularly intriguing in the context of strong AI research. These include their recognized two scientific revolutions, cognitive and probabilistic, in the middle of the twentieth century. Our example here focuses on the latter and its more general form — statistical. Indeed, all is about reasoning with uncertainty and has a long history (Nickerson, 2004). Our brief discussion will focus on inventions for inductive inference that date back to the 260-year-old topic — Bayes (1763).

## 5.1 The Bayesian invention

There have been debates on whether scientific tools such as logic, mathematics, and statistics belong to science. This is particularly relevant because the proposed TB framework we consider here is based on empirical science. Our simple solution is to consider the content for which the statistical method is designed. To be specific, consider the scientific problem of Bayes (1763) from an experiment consisting of $n$ independent and identically distributed Bernoulli trials with the probability of success $\theta \in [0, 1]$. From this experiment, we have observed $X$ successes, $X \leq n$. We want to update our knowledge on $\theta$ or the probability calculations of future such Bernoulli events, assuming the unknown probability of success $\theta$ follows the standard uniform distribution $\text{Uniform}(0, 1)$ *a priori*. The result is the Bayes theorem: given $X$, $\theta$ has the posterior distribution

$$\pi(\theta|X) = \frac{\pi(\theta)f_\theta(X)}{\int_0^1 \pi(\theta)f_\theta(X)d\theta},$$

where $\pi(\theta)$ is the probability density function of the prior distribution Uniform$(0,1)$, $\pi(\theta) = 1$ for all $\theta \in [0,1]$, and $f_\theta(X)$ is the probability mass function of $X$, $\theta^X(1-\theta)^{n-X}$.

Bayes' theorem and Bayesian inference developed over the last 260 years, including the work of Pierr-Simon Laplace (see, *e.g.* Jaynes, 2003; Stigler, 1986) and neo-Bayesian revival (see, *e.g.* Fienberg, 2006), have profoundly shaped mathematics, statistics, and numerous other disciplines and, thereby, gone through a careful process of *TB-exploration* since their inception or *creation*. Nevertheless, while Bayesian remains as a popular school of thought, the dominant school of thought in scientific inference is frequentist, which started to surface a century ago. Here, we take a critical look the Bayesian theory from a perspective of TB-evaluation.

Perhaps, the most acceptable evaluation scheme is to evaluate the predicted future observations from the Bayes model against the outcome that will be actually observed. Mathematically, this ultimately reduces to the *frequentist evaluation* of the posterior distribution when a specific prior distribution $\pi(\theta)$ is used, even though the data analyst has no prior knowledge about the value of $\theta$ *a priori*. The failure of the Bayes theory on frequency evaluation (TB-evaluation), especially on constrained parameter and multi-parameter problems, inspired paradigm shift research for alternative ways of scientific inference.

## 5.2 The fiducial inspiration

Philosophically, the primary foundation of frequentism lies in the interpretation of probability as a measure of long-run frequency of events in repeated trials. This perspective was principally developed (created) by Ronald Fisher, as well as Jerzy Neyman and Egon Pearson. Ronald Fisher made significant contributions by introducing the concept of significance testing, which evaluates how surprising a statistic is with respect to a null hypothesis.

Neyman and Pearson expanded on Fisher's ideas to address scenarios involving multiple competing hypotheses. They proposed that the likelihood ratio, comparing probabilities under different hypotheses, could be used to maximize the differentiation between these hypotheses. Their work introduced the formal framework of Type I and Type II errors, with Type I errors representing false positives (rejecting a true null hypothesis) and Type II errors representing false negatives (failing to reject a false null hypothesis). This framework also established the concept of test power, emphasizing the optimization of tests to balance error probabilities while exceeding a predetermined significance level. Nevertheless, from the TB perspective of refining Bayesian theory that has to be probabilistic to serve as the logic of science, Fisher's concept of significance testing using his invention of p-values is aligned with the logic of science, and his fiducial inference seemed to be on target more than Neyman-Pearson's concepts of confidence of intervals and their framework of Type I and Type II for hypothesis testing. The retrospective discussion of Fisher's p-value in the IM framework is given by Martin and Liu (2014). Here, we discuss fiducial inference with the focus on our TB analysis using the example of Bayes (1763), rather than the simpler, more familiar cases of Fisher (1973).

Since R. A. Fisher didn't develop a complete fiducial theory, fiducial-inspired efforts have appeared in different places (see, *e.g,* Zabell, 1992; Hannig, 2009). The Binomial$(n, \theta)$ model for the observed count $X$ of successes in $n$ iid Bernoulli trials with the probability of success $\theta$ has the probability density function (p.d.f.)

$$f_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad (k = 0, 1, 2, ..., n) \tag{4}$$

and the c.d.f. $F_\theta(x) = 1 - \text{Beta}_{x+1,n-x}(\theta)$, the regularized incomplete beta function $I_{1-\theta}(n-x, x+1)$. Note $\text{Beta}_{a,b}(x) = I_x(a,b)$, $I_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt / \int_0^1 t^{a-1}(1-t)^{b-1}dt$, and $I_x(a,b) = 1 - I_{1-x}(b,a)$. Let $U \sim \text{Uniform}(0,1)$, and define $X$ to be the $U$-th quantile of

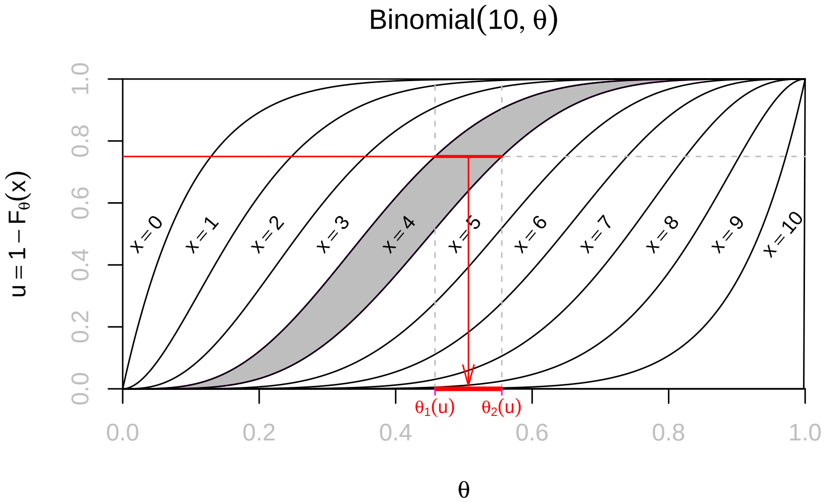


Figure 3: The *fiducial* set-valued mapping $\{\theta : F_\theta(x-1) < 1-u \leq F_\theta(x)\}$ for $u$ given $x$. The gray area is for the case with $n = 10$ and $x = 4$.

$F_\theta$,

$$X = F_\theta^{-1}(U) = \min\{x : U \leq F_\theta(x)\},$$

which serves as a data generating equation with $U$ as what Martin and Liu (2013) call *the auxiliary variable*. Then, we have for $U = u$ and $x = F_\theta^{-1}(u)$

$$\text{Beta}^{-1}_{x,n-x+1}(1-u) < \theta \leq \text{Beta}^{-1}_{x+1,n-x}(1-u). \tag{5}$$

This *fiducial* set-valued mapping, is depicted in Figure 3. For continuous distributions, This *fiducial* set-valued mapping reduces to a usual function. In that case, as shown in Fisher (1973), the fiducial distribution of $\theta$ is obtained by pushing the distribution of $U$ to the space of $\theta$ using this usual function. The discussion of the general set-valued mapping will be continued in Section 5.3.

Unfortunately, R. A. Fisher did not give a general definition of fiducial inference. Most of his examples were for a single parameter, except for the famous Behrens-Fisher problem (see, *e.g.*, Martin and Liu, 2015b); different generalizations have been given when there are several parameters (Tukey, 1957; Stein, 1959). According to Zabell (1992), "the fiducial argument stands as Fisher's one great failure", a sentiment that has been echoed by others (Dawid, 2024), and apparently has a connection with its perception by Jerzy Neyman, one of Fisher's contemporaries and critics. Neyman and Fisher were known to have significant philosophical and methodological disagreements, particularly regarding statistical inference. Neyman, a key proponent of frequentist methods, criticized Fisher's fiducial inference approach as being vague and inconsistent with strict frequentist principles. All these can be viewed as failures with respect to TB-evaluation in terms of frequency evaluation.

R. A. Fisher himself maintained throughout his life that fiducial inference was an important contribution, even if it was not universally accepted or fully developed to his satisfaction. Critics like Neyman often emphasized its logical difficulties and the lack of general applicability, leading to the sentiment that it was a notable misstep in Fisher's otherwise groundbreaking career.

Nevertheless, Fisher's idea of fiducial inference has appeared to be inspirational. It stimulated the tremendous explorations (see, *e.g.* Birnbaum, 1961; Fraser, 1961; Dempster, 1964; Zabell, 1992; Dawid, 2024; Nancy, 2022, and references therein), and continues to inspire creative statistical methods. The latter includes generalized fiducial (Hannig, 2009; Hannig et al., 2016; Liang et al., 2024), which are mostly focused on large-sample-based justification, applications, computational methods, and exact Neyman-Pearson confidence methods (Cook and Weisberg, 1990; Xie and Singh, 2013; Cui and Hannig, 2022; Xie and Wang, 2022).

## 5.3 The Dempster-Shafer discovery

As (the creation of) both a successor fiducial and a generalization of Bayes (see Dempster (2008)), the Dempster-Shafer theory builds upon Dempster's discovery of the need of using upper and lower probabilities for inference and Shafer's development of a broader framework Shafer (1976) to define and manipulate belief functions. Its dual notions of belief and plausibility offer a nuanced way to assess confidence in propositions when full probabilistic information is unavailable.

Here, from the TB perspective, consider the running example of inference with a binomial count. A natural way of extending Fisher's fiducial approach to continuous examples to the discrete cases would lead to considering the set-valued mapping (5):

$$\Theta_x(u) \equiv \{\theta : \text{Beta}^{-1}_{x,n-x+1}(1-u) < \theta \leq \text{Beta}^{-1}_{x+1,n-x}(1-u)\}. \tag{6}$$

Using a predictive interval for the unobserved auxiliary variable $U$, for example,

$$\mathcal{U}_\alpha = \left[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right], \qquad \alpha \in (0,1),$$

we can construct a $(1-\alpha)100\%$ frequentist confidence interval

$$\Theta_x(\alpha) = \cup_{u \in \mathcal{U}_\alpha} \Theta_x(u) = \left[\text{Beta}^{-1}_{x,n-x+1}(\alpha/2),\ \text{Beta}^{-1}_{x+1,n-x}(1-\alpha/2)\right]$$

for $\theta$, which corresponds to the method of Pearson (1920). Interestingly, this supports the perception that frequentist ideas are "in the air" when R. A. Fisher became what we now consider as frequentist.

In addition to the discovery of the necessity of using set-valued inverse mapping and, thereby, lower-and-upper or imprecise probabilities, other innovations of the Dempster-Shafer theory include combining information, Dempster's rule of combination, and the mathematical theory of evidence (Shafer, 1976). From the TB perspective, everything seems to have come together to form a satisfactory logic of science, except for the frequency evaluation that the majority of scientists apparently considered important logically by the nature of science. Viewed as TB-evaluation, this is discussed in depth in Martin et al. (2010) and Zhang and Liu (2011), which eventually lead to the work discussed next in Section 5.4. For an extensive review of the Dempster-Shafer theory and more discussion of its frequency properties, see Yager and Liu (2008), Liu and Martin (2015), Denœux (2016), and Denœux and Li (2018).

## 5.4 The inferential models framework

Creating a fully satisfactory logic of science is probably still an unsolved problem. This was evidently in statistics and science near the turn of this millennium. For example, John W. Tukey (1990) said: "*Today I do not believe there is any chance for a single unifying approach to inference.*" and Bradley Efron (2013) wrote: "*... perhaps the most important unresolved problem in statistical inference is the use of Bayes' theorem in the absence of prior information.*" But, there is an encouraging assessment by a forthcoming review of IMs in *Statistical Science* (Cui and Hannig, 2022). Those authors say:

> *IMs brought a thoroughly novel idea into the foundations of statistics by formalizing a way to assign epistemic probabilities to events that have guaranteed frequentist interpretation. ... They provide a powerful argument for anyone seeking fiducial or objective Bayes distributions on parameter space to consider making calculations on the auxiliary [variables].*

Here, we take a critical look at IMs from the perspective of TB using the running binomial example.

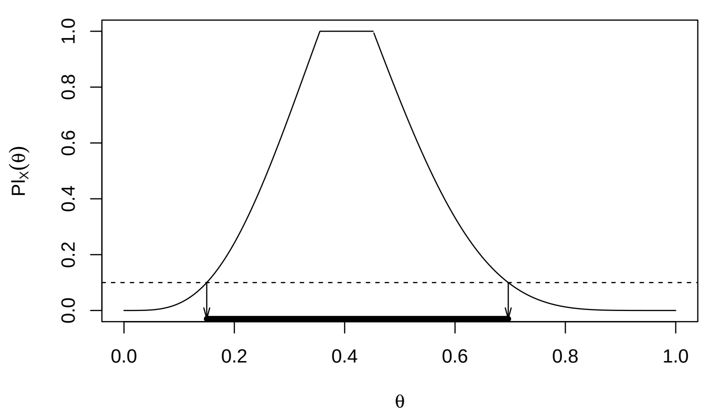


Figure 4: The plausibility curve of the binomial example in Section 5.4 for the case with $n = 10$ and $x = 4$. The 90% plausibility interval is given by the points with $\mathrm{Pl}_X(\theta) \geq 0.1$.

As a computational procedure, IMs consists of three steps, association, prediction, and combination. The association step represents the sampling model by (5), which associates the observed data $X$, the unknown parameter $\theta$, and the auxiliary variable $U$ that is missing but predictable. The prediction step predicts the missing value $U$ using a so-called valid predictive random set $\mathcal{S}$, resulting in *epistemic probabilities*. For example, for a general purpose inference, we can take

$$\mathcal{S} = [U/2, 1 - U/2], \qquad U \sim \text{Uniform}(0, 1),$$

which is valid (Martin and Liu, 2013). The combination step combines the predictive random set $\mathcal{S}$ and the observed data to push $\mathcal{S}$ into a random set on $\Theta$ based on the association (5):

$$\Theta_X(\mathcal{S}) = \cup_{u \in \mathcal{S}} \Theta_X(u),$$

where $\Theta_X(u)$ is defined in (6). For any assertion or hypothesis of interest $A$, this IM model produces the lower probability called *belief*

$$\text{Bel}_X(A) = \text{Prob}(\Theta_X(\mathcal{S}) \subseteq A),$$

the probability of the truth of $A$, and the upper probability called *plausibility*

$$\text{Pl}_X(A) = 1 - \text{Prob}(\Theta_X(\mathcal{S}) \subseteq A^c),$$

the plausibility for the truth of $A$. For example, for all $\theta \in [0, 1]$, we have

$$\begin{aligned} \text{Pl}_X(\{\theta\}) &= 1 - \text{Prob}(\Theta_X(\mathcal{S}) \subseteq \{\xi : 0 \le \xi \le 1, \xi \ne \theta\}) \\ &= \begin{cases} 2\text{Beta}(\theta, X, n - X + 1)), & \text{if } \theta < \text{Beta}^{-1}(0.5, X, n - X + 1); \\ 2(1 - \text{Beta}(\theta, X + 1, n - X)), & \text{if } \theta > \text{Beta}^{-1}(0.5, X + 1, n - X); \\ 1, & \text{otherwise.} \end{cases} \end{aligned} \tag{7}$$

IMs are probabilistic and have desirable frequency properties. In the context of significance testing, the belief and plausibility functions generate (situation-specific) probabilistic

| ***Properties Supporting IMs as a Generalized Logic of Science*** |
|---|
| ✓ Objective and Probabilistic Reasoning |
| ✓ Handling Hypotheses and Uncertainty |
| ✓ Inductive Reasoning |
| ✓ Flexibility Across Domains |
| ✓ Empirical Basis for Hypothesis Testing |
| ✓ Bridging Bayesian and Frequentist Paradigms. |

| ***Challenges and Open Questions*** | ***Future Directions*** |
|---|---|
| ① Scalability | ④ Empirical Applications |
| ② Acceptance in the Scientific Community | ③ Theoretical Development |
| ③ Extensions and Generalizations | ② Educational Efforts |
| ④ Philosophical Foundations | ① Integration with Computational Tools |

Table 1: The LLM-based computational TB-evaluation of IMs as a generalized logic of science.

uncertainty assessments of hypotheses of interest. The IM counterpart of Neyman-Pearson confidence interval is plausibility interval, the collection of singleton assertions with plausibility (7) greater than or equal to the given frequentist error rate $\alpha$. This is illustrated in Figure 4, where the plausibility curve can be viewed as a upside-down confidence curve. It should be noted that, philosophically, unlike Neyman-Pearson's confidence intervals, plausibility intervals are appealing because while guarantee exact frequentist coverage, they offer situation-specific probabilistically interpretable uncertainty assessments. That is, plausibility intervals can be interpreted both freqentistly and Bayesianly.

Since the IMs framework is relatively new, an extensive TB-evaluation depends on future research. Here, we conduct an experiment of computational TB-evaluation using ChatGPT, with a manually dynamic chain of thought and verification approach (Wei et al., 2022; Dhuliawala et al., 2023). The detailed process and intermediate results are provided in Supplementary C. The key concluding results are summarized in Table 1, which we found reasonably meaningful and even valuable for future research, considering that the assessments are done at the language level. Since ChatGPT is generative, we will not over-interpret its results. Nevertheless, formal statistical and computational TB-evaluation is subject to future development.

# 6 Concluding Remarks

Currently, statistics is comprised of a multitude of disparate estimation and hypothesis testing techniques, and when a hypothesis is rejected, it's often unclear how to incorporate the knowledge of that rejection into the new estimation problem. However, in science, the formulation of hypotheses and their testing does not occur in a vacuum, and more plausible alternatives are often required to gain broad traction when rejecting an existing theory. In this work, we introduce an overarching framework that ties together hypothesis testing and estimation for the purposes of making scientific discoveries. This framework provides a direction to reconcile differences in foundational approaches to statistics, such as frequentism, Bayesianism, and inferential models, by placing them in their proper context within the process of creativity. While creation's natural statistical analogue is modeling and estimation and evaluation's is hypothesis testing, there are not yet well established methods for exploration. In our current era of weak-AI development, we have the ability to estimate arbitrary functions and incorporate data of various modalities into prediction machines, but these are all constrained in their capacity by their formulation. Strong AI will need to encompass all three steps of scientific discovery, and improvements in the creation and exploration steps will be necessary to get there.

Technically, we proposed a very simple *statistical* framework for understanding and analyzing creativity (see, *e.g.*, Wiggins, 2006, for more elaborated systems on general creativity). This framework is inspired by a detailed study of significant discoveries in celestial mechanics and physics and is illustrated through both a straightforward artificial example and the extensive 260-year-long pursuit of a sound logic of science. Although we only conducted a limited computational experiment on TB-evaluation of IMs using an LLM with manually dynamic chain of thought and verification, computational TB-creation and TB-exploration can be considered in future research by making use of generative AI models and conducting automodeling, which has motivated relevant research (see, *e.g.*, Sun et al., 2024, and references therein).

However, we acknowledge that TB itself is built primarily on inductive reasoning, which inherently leaves it open to further evaluation and refinement. To enable such evaluations, it would be valuable to collect systematic data on scientific creativity. Such data would not only facilitate rigorous assessment of the proposed framework but also support its expansion, thereby establishing empirical foundations for computational methods to understand and analyze scientific creativity. In the long run, this could lead to methods capable of generating scientific ideas and solving complex problems. At this juncture, it is important to note that exciting progress has already been made in fields such as computational creativity and experimental creativity, where data-driven approaches and experimental observations provide crucial insights (see, *e.g.* Varshney et al., 2019; Soroush et al., 2024). Leveraging these advancements can further strengthen the empirical grounding of our framework and enhance our understanding of the mechanisms behind scientific creativity.

## Acknowledgment

The authors are supported in part by U.S. NSF grant DMS-2412629.

## References

Airy, G. B. (1846). Account of some circumstances historically connected with the discovery of the planet exterior to uranus. Monthly Notices of the Royal Astronomical Society, Vol. 7, p. 121-144, 7:121–144.

Aleinikov, A. G. (2013). Science of Creativity, pages 1574–1592. Springer New York, New York, NY.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, 53(3):370–418.

Bickel, P. J. and Doksum, K. A. (2015). Mathematical statistics: basic ideas and selected topics, volumes I-II package. Chapman and Hall/CRC.

Birnbaum, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. Journal of the American Statistical Association, 56(294):246–249.

Bouvard, A. (1821). Astronomical Tables Published by the Bureau of Longitudes of France: Containing the Tables of Jupiter, Saturn, and Uranus, Constructed According to the Theory of Celestial Mechanics. Bachelier and Huzard.

Carnovalini, F. and Rodà, A. (2020). Computational creativity and music generation systems: An introduction to the state of the art. Frontiers in Artificial Intelligence, 3:14.

Cook, R. D. and Weisberg, S. (1990). Confidence curves in nonlinear regression. Journal of the American Statistical Association, 85(410):544–551.

Cui, Y. and Hannig, J. (2022). Demystifying inferential models: A fiducial perspective. arXiv preprint arXiv:2205.05612.

Dawid, P. (2024). Fiducial inference, then and now. In Handbook of Bayesian, Fiducial, and Frequentist Inference, pages 83–105. Chapman and Hall/CRC.

Dempster, A. P. (1964). On the difficulties inherent in fisher's fiducial argument. Journal of the American Statistical Association, 59(305):56–66.

Dempster, A. P. (2008). The dempster–shafer calculus for statisticians. International Journal of approximate reasoning, 48(2):365–377.

Denœux, T. (2016). 40 years of Dempster-Shafer theory. International Journal of Approximate Reasoning, 79(C):1–6.

Denœux, T. and Li, S. (2018). Frequency-calibrated belief functions: review and new insights. International Journal of Approximate Reasoning, 92:232–254.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., and Weston, J. (2023). Chain-of-verification reduces hallucination in large language models. arXiv preprint arXiv:2309.11495.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. The annals of applied statistics, 6:1971–1997.

Efron, B. (2016). Empirical Bayes deconvolution estimates. Biometrika, 103(1):1–20.

Fienberg, S. E. (2006). When did bayesian inference become “Bayesian”? Bayesian Analysis, 1(1):1–40.

Fisher, R. A. (1973). Statistical methods and scientific inference, Third Edition. Hafner Publishing Co.

Fraser, D. A. (1961). On fiducial inference. The Annals of Mathematical Statistics, 32(3):661–676.

Gelman, A. and Vehtari, A. (2021). What are the most important statistical ideas of the past 50 years? Journal of the American Statistical Association, 116(536):2087–2097.

Gigerenzer, G. and Murray, D. J. (2015). Cognition as intuitive statistics. Psychology Press.

Gregory, D. (1715). The Elements of Astronomy, Physical and Geometrical. By David Gregory... Done Into English, with Additions and Corrections. To which is Annex'd, Dr. Halley's Synopsis of the Astronomy of Comets.., volume 1. J. Nicholson... sold.

Grosser, M. (1964). The search for a planet beyond neptune. Isis, 55(2):163–183.

Hannig, J. (2009). On generalized fiducial inference. Statistica Sinica, 19:491–544.

Hannig, J., Iyer, H., Lai, R. C., and Lee, T. C. (2016). Generalized fiducial inference: A review and new results. Journal of the American Statistical Association, 111(515):1346–1361.

Hanson, N. (1958). Patterns of discovery. cambridge: Cambridge university press.

Hanson, N. R. (2014). Is there a logic of scientific discovery? In Philosophy, Science, and History, pages 220–234. Routledge.

Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK ; New York, NY, annotated edition.

Jiang, Y. and Liu, C. (2024). Estimation of over-parameterized models from an auto-modeling perspective. arXiv preprint arXiv:2206.01824.

Kuhn, T. S. (1970). The Structure of Scientific Revolutions. University of Chicago press Chicago, second edition.

Liang, F., Kim, S., and Sun, Y. (2024). Extended fiducial inference: toward an automated process of statistical inference. Journal of the Royal Statistical Society Series B: Statistical Methodology.

Liu, C. and Martin, R. (2015). Frameworks for prior-free posterior probabilistic inference. Wiley Interdisciplinary Reviews: Computational Statistics, 7(1):77–85.

Martin, R. and Liu, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. Journal of the American Statistical Association, 108(501):301–313.

Martin, R. and Liu, C. (2014). A note on p-values interpreted as plausibilities. Statistica Sinica, pages 1703–1716.

Martin, R. and Liu, C. (2015a). Inferential Models: Reasoning with Uncertainty. Monographs in Statistics and Applied Probability Series. Chapman-Hall/CRC Press.

Martin, R. and Liu, C. (2015b). Marginal inferential models: prior-free probabilistic inference on interest parameters. Journal of the American Statistical Association, 110(512):1621–1631.

Martin, R., Zhang, J., and Liu, C. (2010). Dempster–shafer theory and statistical inference with weak beliefs. Statistical Science, 25(1):72–87.

Nancy, R. (2022). Das fraser: From structural inference to asymptotics. The Canadian Journal of Statistics, 50(4):1104–1115.

Newton, I. (1718). Opticks: or, A treatise of the reflections, refractions, inflections, and colours of light. London: Printed for W. and J. Innys, Printers to the Royal Society.

Nickerson, R. S. (2004). Cognition and chance: The psychology of probabilistic reasoning. Psychology Press.

OpenAI (2024). Chatgpt (december 19 version) [large language model]. https://chat.openai.com/.

Pearson, K. (1920). The fundamental problem of practical statistics. Biometrika, 13(1):1–16.

Popper, K. (2005). The Logic of Scientific Discovery. Taylor and Francis, Abingdon, Oxon, 2nd ed edition. OCLC: 1048597364.

Ratcliffe, S. (2014). Oxford essential quotations. Oxford university press Oxford.

Rodríguez-Moris, G. and Docobo, J. A. (2024). The discovery of neptune revisited. arXiv preprint arXiv:2405.06310.

Schickore, J. (2022). Scientific Discovery. In Zalta, E. N. and Nodelman, U., editors, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, 6(2):461–464. Publisher: Institute of Mathematical Statistics.

Shafer, G. (1976). A mathematical theory of evidence, volume 42. Princeton university press.

Snyder, L. J. (2023). “William Whewell”. The Stanford Encyclopedia of Philosophy (Winter 2023 Edition), Edward N. Zalta  Uri Nodelman (eds.).

Soroush, M. Z., Zhao, M., Jia, W., and Zeng, Y. (2024). Loosely controlled experimental eeg datasets for higher-order cognitions in design and creativity tasks. Data in Brief, 52:109981.

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. The Annals of Mathematical Statistics, 30(4):877–880.

Stigler, S. M. (1986). Laplace’s 1774 memoir on inverse probability. Statistical Science, 1(3):359–363.

Stigler, S. M. (2002). Statistics on the table: The history of statistical concepts and methods. Harvard University Press.

Sun, M., Han, R., Jiang, B., Qi, H., Sun, D., Yuan, Y., and Huang, J. (2024). A survey on large language model-based agents for statistics and data science. arXiv preprint arXiv:2412.14222.

Tukey, J. W. (1957). Some examples with fiducial relevance. The Annals of Mathematical Statistics, pages 687–695.

Varshney, L. R., Pinel, F., Varshney, K. R., Bhattacharjya, D., Schörgendorfer, A., and Chee, Y.-M. (2019). A big data approach to computational creativity: The curious case of chef watson. IBM Journal of Research and Development, 63(1):7–1.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Whewell, W. (1858). Novum organon renovatum. J. W. Parker and son.

Whewell, W. (1996 [1840]). The Philosophy of the Inductive Sciences (Volume II). London: Routledge/Thoemmes.

Wiggins, G. A. (2006). A preliminary framework for description, analysis and comparison of creative systems. Knowledge-based systems, 19(7):449–458.

Xie, M.-G. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. International Statistical Review, 81(1):3–39.

Xie, M.-G. and Wang, P. (2022). Repro samples method for finite-and large-sample inferences. arXiv preprint arXiv:2206.06421.

Yager, R. R. and Liu, L. (2008). Classic works of the Dempster-Shafer theory of belief functions, volume 219. Springer.

Zabell, S. L. (1992). Ra fisher and fiducial argument. Statistical Science, pages 369–387.

Zhang, J. and Liu, C. (2011). Dempster-shafer inference with weak beliefs. Statistica Sinica, 21(2):475–494.