CC BY NC SA
Check for updates

History and Nature of the Jeffreys-Lindley Paradox

Wagenmakers, E.-J., & Ly, A. (2023)

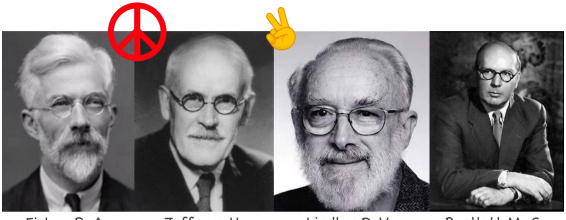
Zhengxiao Wei

April 14 & 21, 2023

Statistical Paradox

The Jeffreys-Lindley paradox refers to the disagreement between the Bayesian and frequentist approaches to a hypothesis testing problem.

Specifically, a p-value (or rejection region or confidence interval) suggests that the point-null hypothesis $\mathcal{H}_0$ should be rejected, whereas the Bayes factor (or posterior probabilities) indicates that $\mathcal{H}_0$ decisively outpredicts the alternative hypothesis $\mathcal{H}_1$.



Fisher, R. A. 1890–1962 | Jeffreys, H. 1891–1989 | Lindley, D. V. 1923–2013 | Bartlett, M. S. 1910–2002

Source: Royal Society, Wikipedia, The Telegraph

Null-Hypothesis Significance Testing

1. Rejection Region: fix α and find the rejection region
$$W = \left(\frac{\hat{\theta} - \theta_0}{\mathrm{se}(\hat{\theta})} \right)^2 \sim \chi^2(1) \tag{1}$$
2. P-Value: compute the p-value of the test statistic
$$p := \mathbb{P}(|T| \geqslant |t_n| \mid \mathcal{H}_0) \tag{2}$$
3. Confidence Interval ("Acceptance Region"): check whether CI for the parameter contains the hypothesized value
$$\mathrm{CI} = \hat{\theta} \pm t^* \times \mathrm{se}(\hat{\theta}) \tag{3}$$

The usual statistical method is to evaluate the observed difference and its standard error, and to say that it is not significant if it is less than a certain **constant** multiple of this error.

Bayes Factor

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \boldsymbol{y})}{p(\mathcal{H}_0 \mid \boldsymbol{y})}}_{\text{posterior odds}} = \underbrace{\frac{p(\boldsymbol{y} \mid \mathcal{H}_1)}{p(\boldsymbol{y} \mid \mathcal{H}_0)}}_{\text{Bayes factor } BF_{10}} \cdot \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{prior odds}} \tag{4}$$

Discrete Data

$$BF_{10} = \frac{\sum_i p(\boldsymbol{y} \mid \mathcal{H}_1, \theta_i) \cdot p(\theta_i \mid \mathcal{H}_1)}{\sum_j p(\boldsymbol{y} \mid \mathcal{H}_0, \theta_j) \cdot p(\theta_j \mid \mathcal{H}_0)} \tag{5.1}$$

Continuous Data

$$BF_{10} = \frac{\int_{\boldsymbol{\theta} \in \boldsymbol{\Theta_1}} p(\boldsymbol{y} \mid \mathcal{H}_1, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta} \mid \mathcal{H}_1) \mathrm{d}\boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in \boldsymbol{\Theta_0}} p(\boldsymbol{y} \mid \mathcal{H}_0, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta} \mid \mathcal{H}_0) \mathrm{d}\boldsymbol{\theta}} \tag{5.2}$$

- **Evidence** (supporting $\mathcal{H}_1$) is anecdotal (BF 1-3), moderate (BF 3-10), strong (BF 10-30), or very strong (BF 30-100).

Bayes Factor (continued)

1. The degree to which the data shift one's prior beliefs about the relative plausibility of $\mathcal{H}_0$ versus $\mathcal{H}_1$.
2. The extent to which data update the prior odds to the posterior odds.
3. The updating factor or the extent to which the data cause revision in belief.
4. The ratio of the probabilities of the data, conditional on the two hypotheses that are being compared.
5. How the prior odds must be changed by the data to become the posterior odds.
6. A function of the probability of the data under the two hypotheses in question.
7. The ratio of the likelihood of one particular hypothesis to the likelihood of another.
8. A likelihood ratio of the marginal likelihood of two competing hypotheses, usually a null and an alternative, both averaged over all possible values of the parameters.
9. The probability of the data given one class divided by the probability of the data given the other class.

Bayes Factor (continued)

1. (Ly et al., 2016, p. 19)

2. (Ly et al., 2016, p. 22)
3. (Rouder et al., 2017, p. 306)
4. (Morey et al., 2016, p. 9)

5. (Morey et al., 2016, p. 12)
6. (Morey et al., 2016, p. 15)
7. (`statisticshowto.com`)
8. (Wikipedia, ChatGPT, new Bing, Bard, Claude, *etc.*)

9. (Chandramouli & Shiffrin, 2016, p. 72)

BF_{10} is smaller than one, indicating that the data are $1/BF_{10} = BF_{01}$ times more likely under $\mathcal{H}_0$ than they are under $\mathcal{H}_1$.

Savage-Dickey Density Ratio

The Savage–Dickey density ratio is a special form of the Bayes factor for **nested** models by dividing the value of the posterior density over the parameters for the alternative model evaluated at the hypothesized value by the prior for the same model evaluated at the same point.

$$BF_{01} = \frac{p(\boldsymbol{y} \mid \mathcal{H}_0)}{p(\boldsymbol{y} \mid \mathcal{H}_1)} = \frac{p(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \mid \boldsymbol{y}, \mathcal{H}_1)}{p(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \mid \mathcal{H}_1)} \tag{6}$$

Bayesian Information Criterion (BIC) Approximation

The BIC for model i with the number of free parameters k_i and the sample size n is

$$\text{BIC}(\mathcal{H}_i) = -2 \ln p(\boldsymbol{y} \mid \mathcal{H}_i, \hat{\boldsymbol{\theta}}_i) + k_i \ln n, \qquad n \gg k_i$$

where $\hat{\boldsymbol{\theta}}_i$ maximizes the likelihood function $\mathcal{L}$.

Note the second-order Taylor approximation of log-marginal likelihood of each model.

$$BF_{01} \approx \exp\left\{\frac{\text{BIC}(\mathcal{H}_1) - \text{BIC}(\mathcal{H}_0)}{2}\right\} \tag{7}$$

Laplace Approximation

Approximate the integrals in Eq. 5.2.

$$\begin{aligned} \underset{\text{multivariate normal}}{p(\boldsymbol{y} \mid \mathcal{H}_i)} &= \int_{\boldsymbol{\theta} \in \boldsymbol{\Theta_i}} p(\boldsymbol{y} \mid \mathcal{H}_i, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta} \mid \mathcal{H}_i) \, \mathrm{d}\boldsymbol{\theta} \\ &\approx (2\pi)^{\frac{k_i}{2}} \cdot |\underset{>0}{\tilde{\mathbf{H}}_i}|^{-\frac{1}{2}} \cdot p(\boldsymbol{y} \mid \mathcal{H}_i, \tilde{\boldsymbol{\theta}}_i) \cdot p(\tilde{\boldsymbol{\theta}}_i \mid \mathcal{H}_i), \end{aligned} \tag{8}$$

where k_i is the number of free parameters ($\dim(\boldsymbol{\Theta}_i) = k_i$),
$\tilde{\boldsymbol{\theta}}_i$ is the maximum *a posteriori* (MAP) estimate (posterior mode),
and $\tilde{\mathbf{H}}_i$ is the Hessian matrix at the MAP estimate.

$$(\mathbf{H}_{\mathcal{L}})_{r,c} = \frac{\partial^2 \mathcal{L}}{\partial \theta_r \partial \theta_c}$$

Bridge Sampling

All terms are conditional on $\mathcal{H}_i$.

$$p(\boldsymbol{y}) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})}[p(\boldsymbol{y} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) \cdot h(\boldsymbol{\theta})]}{\mathbb{E}_{\text{post}}[g(\boldsymbol{\theta}) \cdot h(\boldsymbol{\theta})]} \tag{9}$$

$$\approx \frac{\frac{1}{N_2} \sum_{r=1}^{N_2} p(\boldsymbol{y} \mid \boldsymbol{\theta}_r^{\text{prop}}) \cdot p(\boldsymbol{\theta}_r^{\text{prop}}) \cdot h(\boldsymbol{\theta}_r^{\text{prop}})}{\frac{1}{N_1} \sum_{s=1}^{N_1} g(\boldsymbol{\theta}_s^{\text{post}}) \cdot h(\boldsymbol{\theta}_s^{\text{post}})}, \tag{10}$$

$$\boldsymbol{\theta}_r^{\text{prop}} \sim g(\boldsymbol{\theta}),$$
$$\boldsymbol{\theta}_s^{\text{post}} \sim p(\boldsymbol{\theta} \mid \boldsymbol{y}),$$

where $g(\boldsymbol{\theta})$ is the proposal distribution
and $h(\boldsymbol{\theta})$ is the bridge function.

```
> devtools::install_github("quentingronau/bridgesampling@master")
```

Posterior Model Probability

$$
\begin{aligned}
p(\mathcal{H}_1 \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid \mathcal{H}_1) \cdot p(\mathcal{H}_1)}{p(\boldsymbol{y})} \\
&= \frac{p(\boldsymbol{y} \mid \mathcal{H}_1) \cdot p(\mathcal{H}_1)}{p(\boldsymbol{y} \mid \mathcal{H}_1) \cdot p(\mathcal{H}_1) + p(\boldsymbol{y} \mid \mathcal{H}_0) \cdot p(\mathcal{H}_0)} \\
&= \frac{BF_{10} \cdot p(\mathcal{H}_1)}{BF_{10} \cdot p(\mathcal{H}_1) + p(\mathcal{H}_0)} \\
&= \frac{BF_{10}}{BF_{10} + p(\mathcal{H}_0)/p(\mathcal{H}_1)} \qquad (11)
\end{aligned}
$$

Source

BayesFactor

```
> k <- 49581; N <- 98451; k/N
[1] 0.5036109
> binom.test(k, N)

        Exact binomial test

data:  k and N
number of successes = 49581, number of trials = 98451, p-value = 0.02365
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5004826 0.5067390
sample estimates:
probability of success
             0.5036109

> 1 / BayesFactor::proportionBF(k, N, .5)
Bayes factor analysis
--------------
[1] Null, p=0.5 : 9.607493 ±0%

Against denominator:
  Alternative, p0 = 0.5, r = 0.5, p =/= p0
---
Bayes factor type: BFproportion, logistic

> dbeta(.5, k+1, N-k+1) / dbeta(.5, 1, 1)
[1] 19.21139
```

"Sheldon and Leonard's Whiteboards" (TBBT S08E23)

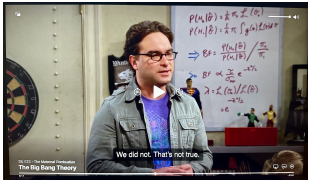
$$p(\mathcal{H}_0 \mid \hat{\theta}) = \frac{1}{A}\,\pi_0\,\mathcal{L}(\theta_0)$$

$$p(\mathcal{H}_1 \mid \hat{\theta}) = \frac{1}{A}\,\pi_1 \int g(\theta)\,\mathcal{L}(\theta)\,\mathrm{d}\theta$$

$$\Longrightarrow BF = \frac{p(\mathcal{H}_0 \mid \hat{\theta})}{p(\mathcal{H}_1 \mid \hat{\theta})} \Bigg/ \frac{\pi_0}{\pi_1}$$

$$\text{So} \quad BF \propto \frac{\tau}{\sigma} e^{-z^2/2}$$

$$\lambda = \mathcal{L}(\theta_0) \;/\; \mathcal{L}(\hat{\theta}) = e^{-z^2/2}$$



Source: Apple TV *Copyright: CBS and Warner Bros.*

Jeffreys Did it First, and Better

The Bayes factor (or the posterior odds) is dependent on **sample size**.

$$BF_{10} \approx A \cdot n^{-\frac{1}{2}} \cdot \exp\{B(\cdot)\}$$

From 1935 to 1939, Jeffreys had repeatedly emphasized the conflict between p-values and Bayes factors. However, his work on Bayes factors had been largely ignored. Instead, it was Lindley (1957) that brought the paradox into the limelight.

Although Lindley's conclusions were qualitatively correct, he did omit a uniform PDF from his equations (thus, overstated), a slip that was corrected by Bartlett (1957).

(p. 49)

Lindley's Paradox

In a normal mean testing problem,

$$\bar{x}_n \sim \mathcal{N}(\theta, \sigma^2/n), \quad \mathcal{H}_0 : \ \theta = \theta_0,$$

under Jeffreys prior, $\theta \sim \mathcal{N}(\theta_0, \sigma^2)$, the Bayes factor

$$BF_{01} = (1+n)^{\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{nt^2}{1+n}\right\},$$

where $t^2 = n(\bar{x}_n - \theta_0)^2/\sigma^2$, satisfies

$$BF_{01} \xrightarrow{n \to +\infty} +\infty,$$

assuming a fixed t^2.

A Fully Frequentist Version of the Paradox

The paradox may be given a purely frequentist interpretation as a discrepancy between (1) minimizing β for fixed α; versus (2) minimizing the weighted sum of errors, $\lambda\alpha + \beta$.

(p. 56-57)

If we keep α fixed as n increases from 20 to 100, we have a rapidly increasing chance of establishing that a difference is significant.

The paradox can be avoided by adopting a lower value of α when power $(1-\beta)$ is known to be high.

(p. 47)

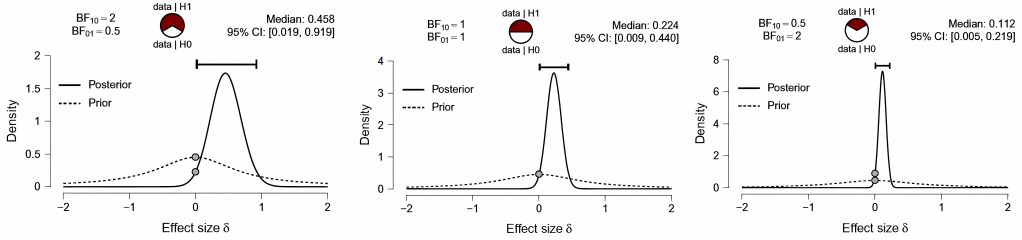
A Fully Bayesian Version of the Paradox

	Two-Sided	Right-Sided	Left-Sided
$\mathcal{H}_0$	$\theta = \theta_0$	$\theta \leqslant \theta_0$	$\theta \geqslant \theta_0$
$\mathcal{H}_1$	$\theta \neq \theta_0$	$\theta > \theta_0$	$\theta < \theta_0$

Data may be constructed which will convince the Bayesian that the population effect is positive rather than negative ... whereas this same Bayesian will also be convinced that the population effect is absent rather than present. ... This state of knowledge is not incoherent, but it may be counter-intuitive.

(p. 57)

A Fully Bayesian Version of the Paradox (continued)



All panels have the same posterior mass on negative effect size and therefore offer the same evidence that the effect is positive rather than negative (i.e., $BF_{+-} = 47.9768$).

At the same time, the test against the point-null hypothesis $\mathcal{H}_0$ reveals increasing support for $\mathcal{H}_0$ as sample size n grows (left: $n = 20$; middle: $n = 82$; right: $n = 332$).

A First Attempt to Escape from the Paradox

Down with Point Masses?

One attempt to question the relevance of the paradox is to argue that the null hypothesis is never true exactly, and it is unwise to assign separate prior mass to a single point from a continuous distribution.

The paradox does not depend on the presence of a point-null hypothesis, as is usually claimed.

Even when the point-null is replaced by a peri-null hypothesis (a relatively peaked continuous distribution), there would be cases, with large numbers of observations, when a new parameter is asserted on evidence that is actually against it.

A Second Attempt to Escape from the Paradox

Blaming the Prior?

Whenever the paradox occurs, a natural objection to the Bayes factor outcome is that the prior distribution for the test-relevant parameter under $\mathcal{H}_1$ was too wide, wasting considerable prior mass on large values of effect size that yield poor predictive performance.

The critique that the prior was too wide is made post-hoc; after observing a near-zero effect size, one may always argue that, in hindsight, the prior was too wide – if such reasoning were allowed, then the data could never undercut $\mathcal{H}_1$ and support $\mathcal{H}_0$.

As long as the prior width does not shrink as a function of sample size, the paradox arises under any non-zero prior width.

Bayes Factor Consistency

An **objective** Bayes factor aims to avoid critical issues:

- Bartlett's paradox

  BF_{10} approaches zero as the prior variance σ^2 increases.

- Information paradox

  BF_{10} tends to be bounded, given overwhelming information t^2.

For the t-test, Gönen et al. (2005) showed the occurrence of these two paradoxes.

$$G\text{-}BF_{10} = \left(\frac{1 + \frac{t^2}{\nu}}{1 + \frac{t^2}{\nu(1+n\sigma^2)}} \right)^{\frac{1+\nu}{2}} (1 + n\sigma^2)^{-\frac{1}{2}}.$$

Remark 4

The paradox is caused by the fact that, as n increases and p-value remains constant, an ever increasing set of parameter values under $\mathcal{H}_1$ is inconsistent with the observed data, decreasing $\mathcal{H}_1$'s average predictive performance (i.e., the marginal likelihood).

If the critical value increases with the sample size suitably fast, then the disagreement between the frequentist and Bayesian approaches becomes negligible as the sample size increases (Naaman, 2016).

Remark 9

The paradox results from the discrepancy between two modes of inference:
(1) evaluating a single model (e.g., fixed-α decision making);
(2) contrasting two models, one of which is relatively simple (e.g., the skeptic's $\mathcal{H}_0$) and one which is more complex (e.g., the proponent's $\mathcal{H}_1$).

- The frequentist finds that $\mathcal{H}_0$ is a poor explanation for the observation (without reference to $\mathcal{H}_1$).
- The Bayesian finds that $\mathcal{H}_0$ is a relatively better explanation for the observation than $\mathcal{H}_1$.

In essence, the apparent disagreement between the methods is not a disagreement at all (**veridical paradox**), but rather two different statements about how the hypotheses relate to the data (Wikipedia).

Model Comparison and Model Selection

A model with the highest Bayes factor in a set of models may nonetheless fit badly.

A model having the highest Bayes factor means nothing more than that the model had the highest amount of evidence in favor of it out of the models currently under consideration.

However, a new model that could be considered may perform substantially better. We have stressed here and elsewhere that a model comparison perspective - as opposed to a model selection perspective - respects the fact that the evidence is always **relative** (Morey et al., 2016).

The aim of the Bayes factor is to quantify the support for a model over another, regardless of whether these models are **correct** (Ly et al., 2016).

Case Study: Prosecutor's Fallacy

> *Sally Clark was an English solicitor who, in November 1999, became the victim of a miscarriage of justice when she was found guilty of the murder of her two infant sons. Clark's first son died in December 1996 within a few weeks of his birth. Her second son died in similar circumstances in January 1998.*

Although the data may be unlikely under $\mathcal{H}_0$, they are even less likely under $\mathcal{H}_1$.

$\mathcal{H}_0$: Sally Clark was innocent.
It may be unlikely that two infants both died naturally.
It may be also very unlikely that a mother killed her two children.

Contribution

- Irrevocably, refuted the common misconception that Jeffreys had ignored or neglected the paradox.
- Supported the claim that the paradox in fact presents a defining feature of the Bayes factor hypothesis test.
- Demonstrated the different ways in which Jeffreys explained why a measure of evidence cannot depend on a constant multiple of the standard error.

How about the Credible Interval?

Recall *the first attempt to escape from the paradox.*

Based on the asymptotic behaviour of the posterior, one might reasonably conclude that, with a sufficiently large sample size, the model-averaged **credible interval** will approximate the frequentist's confidence interval for any $(1 - \alpha) \times 100\%$. However, Wagenmakers and Ly (2021, p. 62) argue that "the Jeffreys-Lindley paradox still applies".

Wei, Nathoo, and Masson (2023) investigated the functional relationship between the two quantities.

In this scenario, model selection (i.e., evaluating the relative values of $p(\mathcal{H}_0 \mid \boldsymbol{y})$ and $p(\mathcal{H}_1 \mid \boldsymbol{y})$) is **not** addressing the same question as estimation (i.e., evaluating $p(\theta \mid \boldsymbol{y})$ to determine which values of θ are *a posteriori* most likely)

(Campbell & Gustafson, 2022).

Two Approaches in Parameter Estimation

Frequentist

- Probability is "long-run frequency of repeated events"
- $\mathbb{P}(X \mid \theta)$
  is a sampling distribution
  - function of X with θ fixed
  - Assuming repetitive sampling from the population to find the single true value for a parameter governing the population
- Null-hypothesis significance testing (NHST) and p-value
- Confidence interval

Bayesian

- Probability is "degree of certainty about values"
- $\mathbb{P}(X \mid \theta)$
  is a likelihood
  - function of θ with X fixed
  - Assuming the probability distribution of a parameter value and its reliability is increased by increasing the sample size
- Prior (**subjective**?)
- Posterior $\propto$
  Prior $\times$ Likelihood
- Credible interval

Confidence Interval and Credible Interval

Confidence Interval

parameters are fixed but unknown and data are random;

the lower and upper bounds are random

A 95% confidence interval means that with a large number of repeated samples, 95% of such calculated confidence intervals would include the true value of the parameter.

Credible Interval

parameters are random and data are fixed;

the lower and upper bounds are fixed

A 95% credible interval is an interval within which an unobserved parameter value falls with a 95% probability.

References

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*, 533–534. doi:10.1093/biomet/44.3-4.533.

Campbell, H., & Gustafson, P. (2022). Defining a credible interval is not always possible with "point-null" priors: A lesser-known consequence of the Jeffreys-Lindley paradox. *ArXiv*. 1-15. doi:10.48550/arXiv.2210.00029.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80-97. doi: 10.1016/j.jmp.2017.09.005

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*, 203-222. doi:10.1017/S030500410001330X.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192. doi:10.2307/2333251.

Naaman, M. (2016). Almost sure hypothesis testing and a resolution of the Jeffreys-Lindley paradox. *Electronic Journal of Statistics*, *10*, 1526-1550, doi:10.1214/16-EJS1146.

Robert, C. P. (2013). *Bayes* 250^{th} *versus Bayes 2.5.0*. European Meeting of Statisticians, Budapest, Hungary.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804. doi:10.3758/BF03194105.

Wagenmakers, E.-J. (2022). Approximate objective Bayes factors from p-values and sample size: The $3p\sqrt{n}$ rule. *PsyArXiv*. 1-50. doi:10.31234/osf.io/egydq.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*, 158-189. doi:10.1016/j.cogpsych.2009.12.001.

Wagenmakers, E.-J., & Ly, A. (2022). Everything you always wanted to know about the Jeffreys-Lindley paradox but were afraid to ask. *Bayesian Spectacles*.

Wagenmakers, E.-J., & Ly, A. (2023). History and nature of the Jeffreys-Lindley paradox. *Archive for History of Exact Sciences*, *77*, 25-72. doi:10.1007/s00407-022-00298-3.

Wei, Z., Nathoo, F. S., & Masson, M. E. J. (2023). Investigating the relationship between the Bayes factor and the separation of credible intervals. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-023-02295-1.

The screenshots from "The Big Bang Theory: The Maternal Combustion" (2015) are used for educational commentary purposes and are believed to be covered by fair use provisions.

Thank you.

Contact: zhengxiao@uvic.ca