Robust and Replicable Bayesian Process Tracing: An Application to Dissents in the United States Supreme Court

Preliminary Version: This paper is under active development. Results and conclusions may change as research progresses. *

Justin Esarey *Wake Forest University*
Susan Smelcer *Wake Forest University*

Bayesian process tracing is a powerful technique for integrating multiple pieces of qualitative and quantitative evidence to reach a conclusion that is logically, transparently, and reproducibly connected to that evidence. We believe it should be a core tool for empirical political scientists. But there is a steep learning curve, exacerbated by methodological disagreements concerning technique and interpretation. We introduce software that facilitates more widespread use of Bayesian process tracing. Our software facilitates inference for multiple paradigms of process tracing. It can compare a theory to its logical complement (e.g., the null), a mutually exclusive and non-exhaustive rival, or a partially rival or complementary theory. We demonstrate our techniques with a qualitative causal analysis of why dissents became much more common in the United States Supreme Court in the early twentieth cenury.

Keywords: process tracing, Bayesian methods, replicability

Bayesian process tracing (BPT) is a powerful technique for integrating multiple pieces of evidence to reach a conclusion that is logically connected to that evidence (Rohlfing, 2012, Chapter 8; Bennett, 2015; Fairfield and Charman, 2022; Humphreys and Jacobs, 2023; Waldner, 2026). It can integrate qualitative and qualitative evidence into a single analysis. It enables a researcher to transparently and reproducibly communicate their interpretation of the evidence (Elman and Kapiszewski, 2014; Moravcsik, 2014; Elman, Gerring and Mahoney, 2016), and in turn to examine the sensitivity of their conclusions to that interpretation. For all these reasons, we believe it should be a core tool for empirical political scientists.

Despite these advantages, BPT has not been widely adopted in the discipline. We think there are two primary reasons, both of which we aim to overcome with this paper.

*__Current version__: November 09, 2026; __Corresponding author__: justin@justinesarey.com.

The first is that "substantial time, effort, and training [is] required" to employ Bayesian methods (Bennett, Fairfield and Soifer, 2019, p. 11). Applying these frameworks can be "a very tall order" because it is not a standard part of current qualitative research practices: "rarely do scholars indicate probabilities for these observations for different possible causal effects or specify the uncertainty they hold over these probabilities" (Humphreys and Jacobs, 2015, p. 671). Explicit, mathematical Bayesian reasoning is an important part of BPT because "linking evidence to inference lies at the heart of analytic transparency" (Bennett, Fairfield and Soifer, 2019, p. 2). It "ensures clear identification and careful assessment of salient evidence" and "facilitates more effective communication of degrees of belief" (Fairfield and Charman, 2017, p. 376). It pushes "researchers to make specific and public the assumptions they must make implicitly for process tracing to work" (Bennett, 2015, p. 297). Unfortunately, "a day or two of intensive workshops is not adequate to successfully apply this approach" (Fairfield and Charman, 2017, p. 376). Given that the "start-up costs are significant" for using them, BPT must "justify the opportunity costs of investing in it" (Zaks, 2021, p. 71).

The second reason is that BPT scholarship disagrees on how evidence should be translated into conclusions. For example, some sources recommend assigning prior beliefs on the log odds of two mutually exclusive theories (Fairfield and Charman, 2022, pp. 83-95), then using Bayes' rule to update these log odds based on the relative consistency of the theories with observed facts. Others argue for comparing the patterns of possible facts predicted by a single model (represented by a relatively simple directed acyclic graph) to clues (pieces of evidence) that the researcher observes in order to update the posterior probability of counterfactual causal relationships depending on their (in)consistency with the observed clues (Humphreys and Jacobs, 2023, pp. 116-130, 155-188). Still others recommend building a directed acyclic graph model of causality based on observed relationships (Waldner, 2026). None of these techniques allows the comparative assessment of *partially rival* and *partially complementary hypotheses* from theories whose truth statuses

are neither completely mutually exclusive nor completely unrelated (Zaks, 2021, p. 67):

> The Bayesian approach is not equipped to handle the range of forms hypotheses and evidence tend to take. The problems are both substantive and mathematical, which together result in a method that is more limited than any of its proponents acknowledge. Substantively, the method encourages an oversimplification of the world by sidestepping the frequency with which two causal factors together bring about an outcome. The upshot is that "[i]n the social sciences, there is no clear procedure for translating complex, narrative-based, nonreproducible, often qualitative information into precise probability statements" (Fairfield and Charman, 2017, p. 376).

Thus, "although it appears to provide a rigorous template for iterative research, methodologists implementing the technique exhibit contradictory and counterintuitive practices when it comes to updating priors as they examine additional evidence" (Zaks, 2021, p. 71). There is also a need to "test whether and to what extent biases arise in practice when conjuring and evaluating quantities like the prior and likelihood functions" (p. 72).

In this paper, we present new software designed to facilitate Bayesian process tracing by scholars without extensive technical training. The software is freely available on the web, usable by anyone with a browser and an internet connection, and its source code can be downloaded to run locally if a researcher desires. The software provides a standardized format for specifying theories, entering descriptions of and/or direct links to qualitative or quantitative evidence, and visually specifying the consistency of the evidence with these theories using graphical illustrations. It is flexible enough to allow analysis under a variety of frameworks for process tracing. It also implements a new method—developed and presented in this paper—for using evidence to simultaneously update our posterior beliefs about two theoretical explanations that are not mutually exclusive but whose truth status may be interdependent. Our software has multiple advantages:

1. It **converts researcher beliefs about theory and evidence into a rigorous form using graphical illustrations**, and then uses those beliefs to draw conclusions with Bayes' Rule that are also graphically depicted. This facilitates the most challenging part of BPT: converting qualitative interpretations into a mathematical version compatible with Bayes' Rule. We demonstrate this conversion both in the toy example and the more complex application that follows.
2. It **allows for easily testing the robustness of conclusions to different intepretations of evidence** and possible biases that might have been present in those interpretations. All an investigator needs to do is to adjust the inputs given to the software.
3. It **accommodates multiple approaches to process tracing**, including the comparison of mutually exclusive and exhaustive theories using log odds (Fairfield and Charman, 2022) but also the comparison of a theory to its logical complement (e.g., the "null hypothesis" of the theory being false) or two theories whose truth status is neither mutually exclusive nor mutually exhaustive.
4. The software **allows researchers to save, share, and revisit earlier analyses**.
5. The **source code is open under a Creative Commons license** for those who wish to examine, modify, or incorporate it into other software.

In the following sections, we first describe how to use our software with a simplified toy example taken from the literature. We next present our new approach to drawing inferences about theories that are not mutually exclusive. Finally, we present an analysis of a more substantive and detailed example: a qualitative investigation of why the number of published dissents issued by the United States Supreme Court grew dramatically during the early twentieth century.

Introducing our Bayesian Process Tracing software

Our web application, available at https://shiny.justinesarey.com/bpt-app, allows users to perform BPT without coding or complex mathematics. It uses the Shiny Server web hosting platform (Chang et al., 2026), which relies on R for the underlying computation and data visualization. The underlying code for the Shiny application is available as part of the replication materials for this paper so that a user can run it directly on their computer as long as R, RStudio, and the necessarily component libraries are installed. But the user need not run this code directly: all they need to do is access the application through an ordinary web browser.

Throughout this section, we will demonstrate our techniques using a canonical[1] toy example of BPT: the investigation of a crime. This example does *not* suffice as a detailed demonstration of the complexity of using process tracing for scientific inference, but it *is* effective as an easy to understand illustration of how to use our software; our study of Supreme Court dissents later in the paper will demonstrate how to handle a more complex and realistic example.

The interface

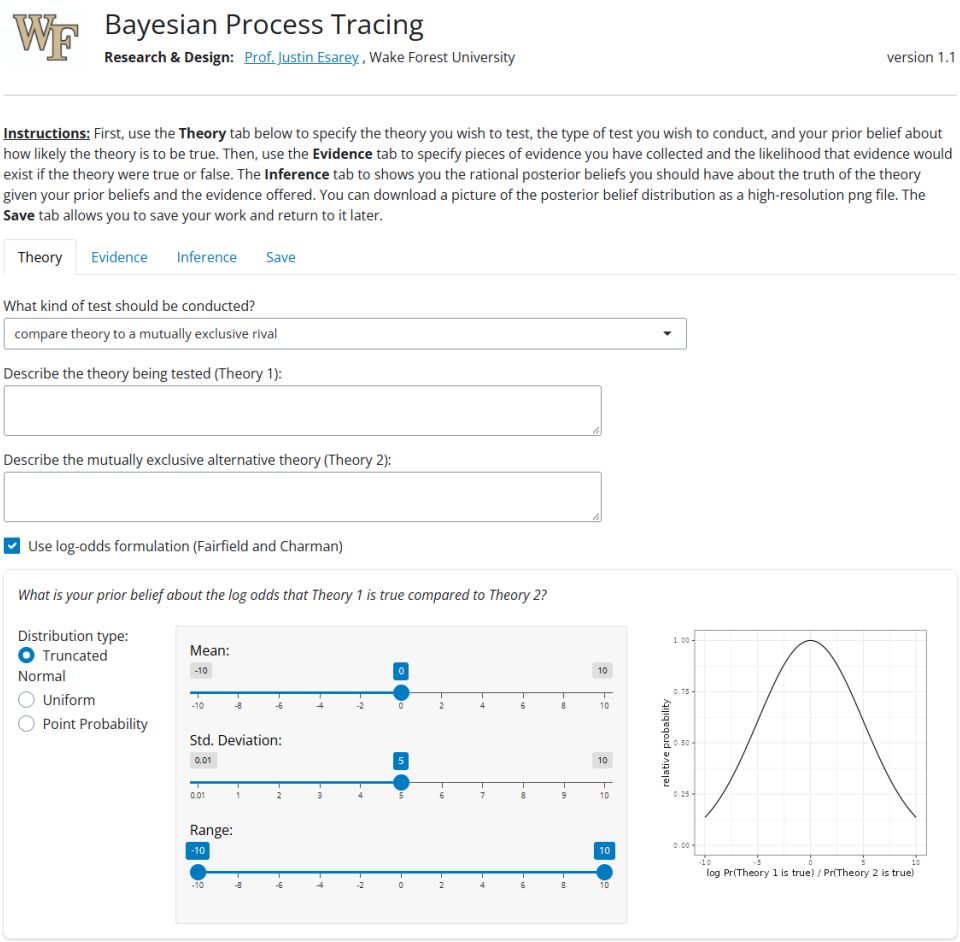
A screenshot of the Shiny application on initial loading is shown in Figure 1. The initial **Theory** tab allows the user to input the nature of the theory (including any alternative, if necessary) and to specify prior beliefs (including interdependence among rival theories, where needed). The application allows for three different types of test to be conducted:

1. the default, comparing a theory to a mutually exclusive and non-exhaustive rival hypothesis as recommended by Fairfield and Charman (2022), with options to input prior beliefs and interpretation of evidence using log-odds or probabilities;

[1] See, *inter alia*, Collier (2011).

2. comparing a theory to its logical complement (the theory being false, sometimes referred to as the null hypothesis); and
3. simultaneously assessing two partially rival, non-exclusive, and non-exhaustive theories with a possibly interdependent truth status.

Figure 1: A screenshot of the initial screen of the Bayesian Process Tracing Shiny app



Use of the application proceeds using the tabs at the top of the screen. After the theory and prior beliefs are specified, evidence is entered in using the second **Evidence** tab. Multiple pieces of evidence can be entered here. For each piece of evidence, a user can specify how consistent the evidence is with the theory being true or false. Once this is done, the application will produce the posterior belief distribution corresponding to the tested theory in the **Inference** tab. The final **Save** tab allows a researcher to save and share

the results of their analysis.

Probably the most challenging aspect of Bayesian Process Tracing is converting a researcher's interpretation of the prior probability for a theory or the consistency of evidence with that theory into a mathematical format amenable to inference using Bayes' Rule. We illustrate how our software facilitates this process with a toy example.

A toy example: Investigating a Crime

As in Collier (2011), we show how our software works using the example of a criminal investigation for which we have identified a pool of potential suspects. Consider the potential guilt of Suspect 1; we may be able to conduct a "hoop test" of their guilt (Van Evera, 1997, p. 31) by checking whether he or she has an alibi, which—if it exists—demonstrates that a suspect could not have committed a crime. Of course, the *absence* of an alibi does not demonstrate that the suspect is guilty, although it may raise the belief of the suspect's guilt. In words, a hoop test uses evidence that cannot exist if the theory is false and might exist if it is true.

This qualitative interpretation can be translated into the language of mathematical Bayesian inference. For a hoop test of a theory T which is either true or false ($\neg T$), we look for evidence x. We then update our belief in the theory accordingly. For a hoop test using binary evidence $x \in \{x_0, x_1\}$, such as in the example of having an alibi or not, then:

$$\Pr(x = x_0|T) = p = 1 \text{ and } \Pr(x = x_1|T) = 1 - p = 0$$

That is, a guilty suspect will never have an alibi. But an innocent suspect may or may not have an alibi; plenty of innocent people cannot account for their whereabouts at all times. Thus:

$$\Pr(x = x_0|\neg T) = q \in [0, 1] \text{ and } \Pr(x = x_1|\neg T) = 1 - q$$

We can then use this evidence along with our prior (pre-evidence) belief that the theory

is true, $\Pr(T)$, to rationally update our beliefs on the basis of the evidence we observe:

$$\begin{aligned}\Pr(T|x) &= \frac{\Pr(x|T)\Pr(T)}{\Pr(x)} \\ &= \frac{\Pr(x|T)\Pr(T)}{\Pr(x|T)\Pr(T) + \Pr(x|\neg T)(1-\Pr(T))}\end{aligned}$$

To continue our previous example, when a suspected perpetrator of a crime has an alibi, $x = x_1$, we conclude that the theory they are guilty (T) is false because $\Pr(x = x_1|T) = 0$ and therefore the numerator of $\Pr(T|x = x_1)$ must be zero:

$$\Pr(T|x = x_1) = \frac{\Pr(x = x_1|T)\Pr(T)}{\Pr(x = x_1|T)\Pr(T) + \Pr(x = x_1|\neg T)(1-\Pr(T))} \tag{1}$$

$$= \frac{(1-p)\Pr(T)}{(1-p)\Pr(T) + (1-q)(1-\Pr(T))} = 0 \tag{2}$$

Formal mathematical reasoning is hardly necessary to draw this conclusion: when a hoop test is failed, this leads to unequivocal rejection of the theory. But when a hoop test is passed (when the suspect does not have an alibi, $x = x_0$, in our running example) we still learn something. It's here where we begin to see the utility of translating the logic of process tracing into Bayesian language:

$$\begin{aligned}\Pr(T|x = x_0) &= \frac{\Pr(x = x_0|T)\Pr(T)}{\Pr(x = x_0|T)\Pr(T) + \Pr(x = x_0|\neg T)(1-\Pr(T))} \\ &= \frac{p\Pr(T)}{p\Pr(T) + q(1-\Pr(T))} \\ &= \frac{\Pr(T)}{\Pr(T) + q(1-\Pr(T))}\end{aligned} \tag{3}$$

Our belief about T, that a particular suspect is guilty of a crime, now depends on three factors: how often guilty people do not have alibis, $p = \Pr(x = x_0|T)$ (which, in a hoop test, must be equal to 1); how often innocent people do not have alibis, $q = Pr(x = x_0|\neg T)$; and our *a priori* belief that the suspect was guilty, $\Pr(T)$. If an investigator wanted to know how they should change their belief about a suspect's guilt when the suspect had

no alibi, they could draw that conclusion by specifying these quantities. The conclusion flows mathematically from them, and that conclusion can be verified and reproduced by specifying them. Furthermore, the sensitivity of the conclusion to these quantities can be assessed by systematically varying them.

Translating evidence into probabilities

Now, suppose that we find that Suspect 1 has no alibi. How do we interpret this evidence and draw a conclusion about his or her guilt? As noted above, translating qualitative interpretations of evidence into a concrete form is the most challenging task of BPT. We must choose:

1. p: *What proportion of guilty suspects have no alibi?* This must be close to 100%, as only in very unusual circumstances would a person responsible for a crime be able to apparently prove they were somewhere else when the crime was committed. Perhaps there is a small chance that a guilty person has an alibi, perhaps because they have been able to falsify one or because of a case of mistaken identity. Thus, we set p at 95% and $(1 - p)$ at 5%. We can test the sensitivity of our conclusion to this interpretation of the evidence by changing it later.
2. q: *What proportion of innocent suspects have no alibi?* Anyone (except the guilty party) who cannot establish their whereabouts during the time frame of the crime fits into this category. Thus, this is probably a large number. Thus, we begin setting $q = (1 - q) = 50\%$, and once again we can assess the sensitivity of our conclusion to this figure later.
3. $\Pr(T)$: *What is our a priori belief that the suspect is guilty?* This is how likely we thought the suspect was to be guilty before examining any new evidence, perhaps based on background information about the suspect's prior history or character. We will begin by specifying a 50% prior probability that the theory is true.

We inputted these details into our BPT software, as shown in Figure 2. and saved the results at https://tinyurl.com/7a7res2a.

Under these circumstances, we have a weak belief that Suspect 1 is guilty: $\Pr(T|x = x_0)$ is 65.5%. Thus, we see that while the suspect lacking an alibi is suggestive, it is hardly decisive.

Robustness assessment and uncertain interpretation of evidence

How much does this conclusion depend on the interpretation of the evidence and our prior beliefs? Suppose, for example, we overestimated the proportion of innocent suspects without an alibi. To see how our conclusion changes when we change this interpretation of the evidence, we can adjust q downward from 50% to 40%, 30%, and 20%. This changes our conclusion about Suspect 1's guilt from 65.5% to 70.4%, 76.0%, and 82.6% respectively. Thus, even an extremely generous interpretation of the evidence is not enough to clinch this suspect's guilt.

However, we may not have enough confidence in our assessment of the evidence to assign a fixed value to the quantities in equation 3. For example, our qualitative assessment of the probability innocent people do not have alibis $\Pr(x = x_0|\neg T)$ is quite uncertain, as the relevant population of innocent suspects is somewhat ill-defined and our knowledge of that population is vague. In that case, we need to represent our uncertainty by specifying a probability distribution $g(q)$ that represents the possible values of q and how likely we think they are to be true. If we have no idea how often innocent people have alibis, we can represent this ignorance by setting $g(q) \sim U[0,1]$, a uniform density that assigns equal probability to every value of q between 0 and 1. This distribution is shown in Figure 3.

Then, we need to integrate over $g(q)$ in the formula for Bayes' Rule:

$$\Pr(T|x = x_0) = \int \frac{p \Pr(T)}{p \Pr(T) + q(1 - \Pr(T))} g(q) dq$$

Figure 2: An lacking alibi as evidence of guilt

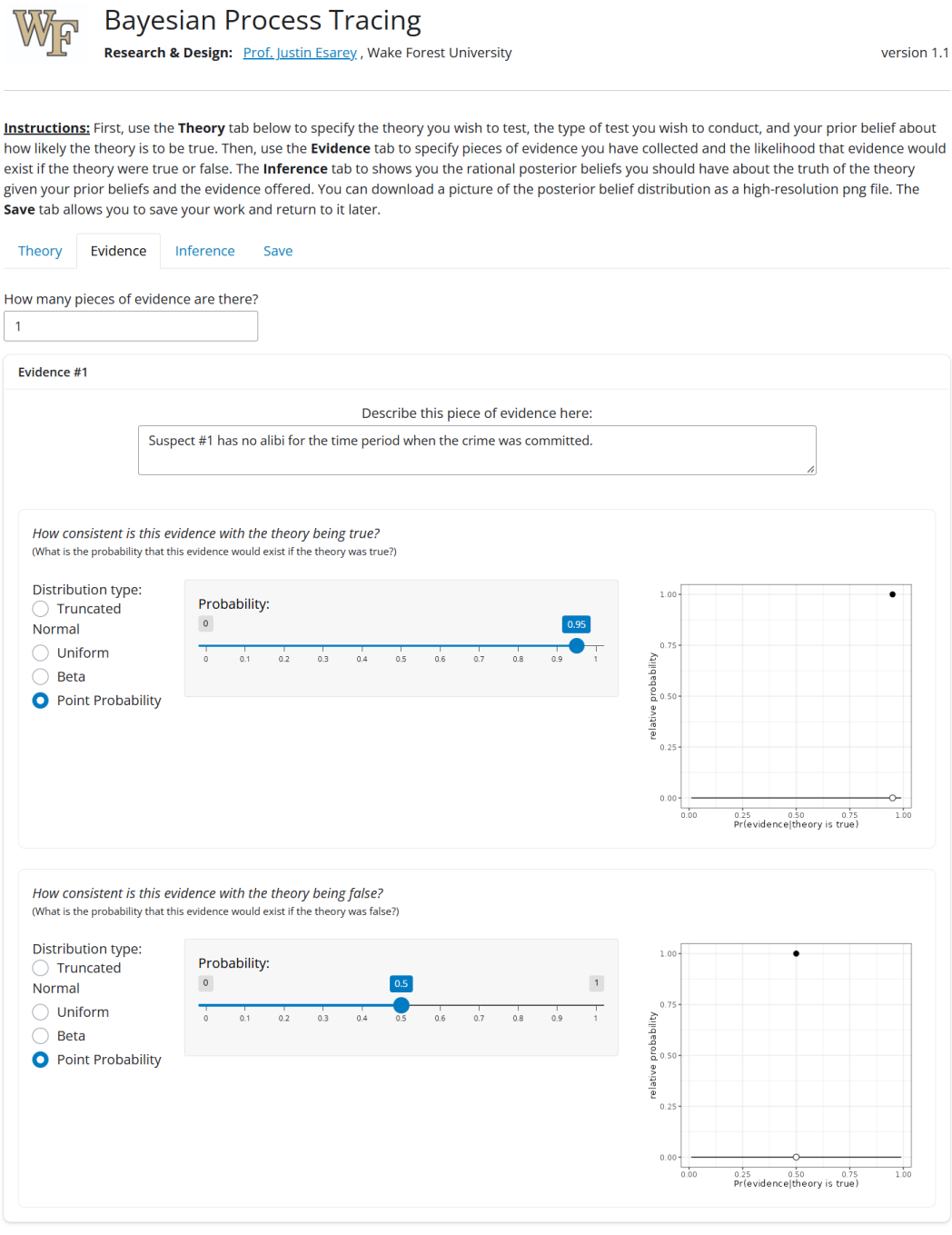
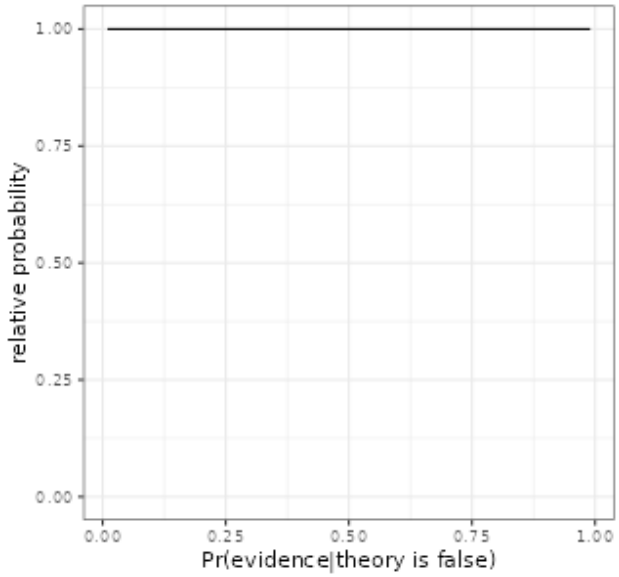


Figure 3: Complete uncertainty about how often innocent suspects have alibis



When there are multiple uncertain terms in the calculation, they must all be integrated out. Continuing the above example, we may also be uncertain about p, the probability that guilty people do not have an alibi. In this case, we must determine:

$$\Pr(T|x = x_0) = \iint \left(\frac{p \Pr(T)}{p \Pr(T) + q(1 - \Pr(T))} f(p|q) \right) g(q) \, dp \, dq \tag{4}$$

This is our posterior belief about the probability that the suspect is guilty. However, it is important to point out that $\Pr(T|x = x_0)$ is *itself* uncertain; there is an implicit distribution $h(t)$ which gives the probability (density) that $\Pr(T|x = x_0)$ equals some candidate value $t \in [0, 1]$. For a particular value of t, this is:

$$h(t|x = x_0) = \iint_{S(t)} f(p|q) g(q) \, dp \, dx$$
$$S(t) = \left\{ (p, q) \middle| \frac{p \Pr(T)}{p \Pr(T) + q(1 - \Pr(T))} = t \right\}$$

Thus, in principle, we can calculate not only our expectation of the probability that the theory is true. We can also determine how certain we are about that belief, and express this uncertainty precisely as a probability density function $h(t|x = x_0)$ that is rigorously

and reproducibly linked to the evidence we laid out in advance.

As the number of uncertain terms expands, these calculations become arbitrarily more analytically complex. However, this is a common and surmountable problem in Bayesian statistics: we can use techniques from Monte Carlo integration instead of analytical calculation to accomplish them all (Robert and Casella, 2004). In the example above, as long as the distributions $f(\bullet)$ and $g(\bullet)$ are known, we can draw many samples from them and calculate the value of $\Pr(T|x = x_0)$ for each sample. The result is a set of samples of values of t from the posterior belief distribution $h(t|x = x_0)$. The expected value in equation 4 can be calculated by simply calculating the mean of these samples, with the accuracy of this calculation an increasing function of the number of samples of t we create. Our software executes this procedure and returns summary statistics and visualizations of the resulting posterior belief distribution.

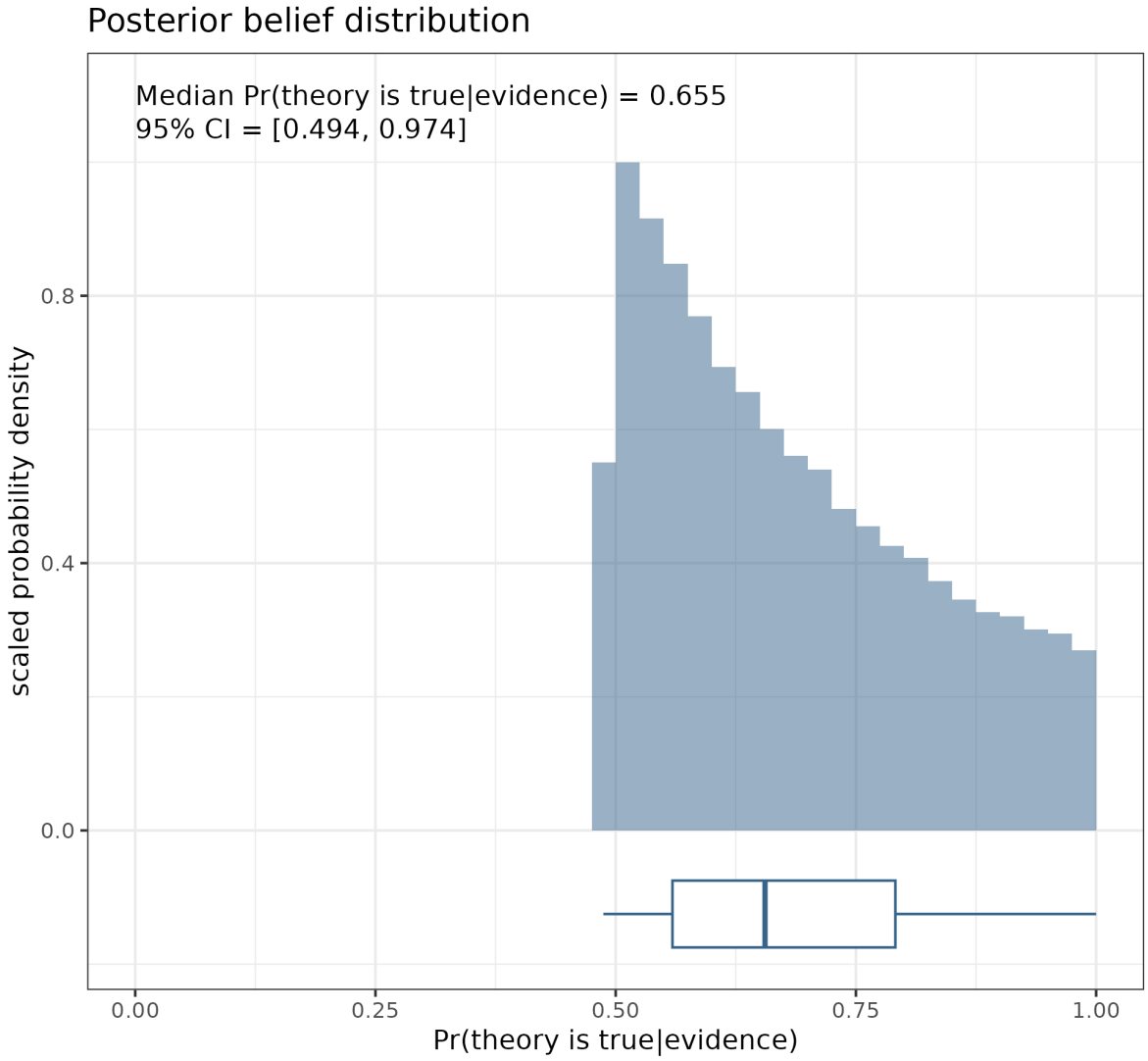
Returning to our example of a suspect with no alibi, when we allow $g(q) \sim U[0, 1]$ the resulting posterior belief about the suspect's guilt shown in Figure 4. As the figure shows, our expectation that the suspect is guilty is virtually unchanged (65.5%). However, there is considerable uncertainty about this probability: we believe that there is a 50% chance that the suspect is actually less likely to be guilty, perhaps as low as a coin flip.

Compound evidence

The example above considers a single piece of binary evidence. Although formal analysis is useful in this case, it is among the simplest qualitative inference scenarios possible. Our analysis is much more useful when there are many pieces of evidence, that evidence does not point directly to a single conclusion, and our beliefs about that evidence are uncertain. Consider evidence $y \in \{y_0, y_1\}$:

$$\Pr(T|y = y_0) < \Pr(T) < \Pr(T|y = y_1)$$

Figure 4: Updated inference about the probability that Suspect 1 is guilty



To continue the example from before, y is akin to whether the suspect has a *motive* to commit a crime. Motive does not imply guilt, and a lack of motive does not imply innocence. However, we would expect that discovering that a person has a motive would raise our assessment of the probability of their guilt, while discovering they have no motive would reduce that assessment. And, if a person had no alibi and did have a motive, the combination would be more incriminating than either in isolation. In the case where we have already provided p_x and q_x from equation 3, we need only provide the additional quantities p_y and q_y to determine the updated belief $\Pr(T|x = x_0, y = y_1)$:

$$\begin{aligned}\Pr(T|x = x_0, y = y_1) &= \frac{\Pr(x = x_0, y = y_1|T)\Pr(T)}{\Pr(x = x_0, y = y_1|T) + \Pr(x = x_0, y = y_1|\neg T)} \\ &= \frac{p_x p_y \Pr(T)}{p_x p_y \Pr(T) + q_x q_y (1 - \Pr(T))}\end{aligned}$$

$$p_x = \Pr(x = x_0|T) \qquad q_x = \Pr(x = x_0|\neg T)$$

$$p_y = \Pr(y = y_1|T, x = x_0) \qquad q_y = \Pr(y = y_1|\neg T, x = x_0)$$

Now suppose that Suspect #1 from our previous example has no alibi and also has a motive to commit the crime. To form a belief about their guilt, in addition to the information from the prior example we must also specify:

1. p_y: *What proportion of guilty suspects have a motive?* Some types of criminals, such as serial killers, may commit crimes in the absence of some financial or emotional connection to the victim. However, we think it it safe to presume that most crimes are committed for a reason. We set this at 90%.
2. q_y: *What proportion of plausible but innocent suspects have a motive?* The reference group is critical here. The vast majority of people in the full population have no motive; indeed, they have no connection at all to anything related to the crime. But for most of these people our prior probability of guilt would also be ≈ 0. These cases represent zero-weight components in the denominator of $\Pr(T|x, y)$. Thus, $\neg T$ should be understood to encompass a set of alternative theories, $(\mathbb{Q} \setminus T) \subseteq \Omega$

such that $\Pr(Q \in \mathbb{Q}) > 0$ is the set of *plausible* alternatives with non-zero prior probability. Even with this limiting proviso, the answer may vary widely according to the situation; victims can have few people who wish them harm, or many. For the moment, we will set this at 50% and revisit that choice later.

We add this information to our previous analysis and save the results at https://tinyurl.com/439j39wt. Our belief in Suspect #1's guilt rises to 77.4% when they have both motive and opportunity to commit the crime.

Robustness assessment and uncertain interpretation of evidence

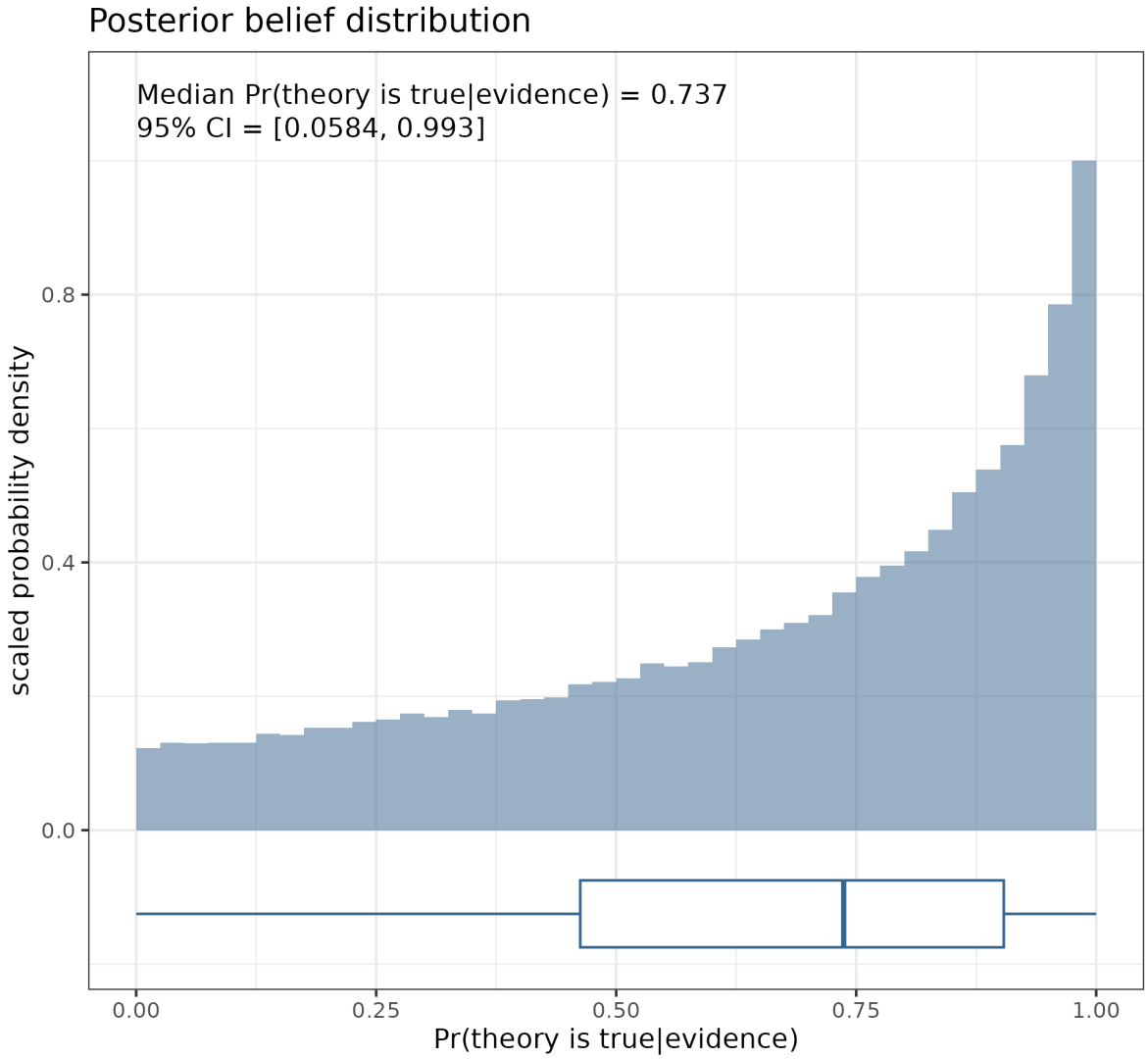
As before, we are uncomfortable assigning fixed probabilities to our interpretation of the evidence and the prior probability that the suspect is guilty. As laid out above, there are five pieces of information that must be supplied and about which we might be uncertain: p_x, q_x, p_y, q_y, and the prior $\Pr(T)$. As probabilities, these are bounded between 0 and 1 and any distributions must be similarly bounded. We use truncated normal distributions $\phi_T(\bullet)$ and uniform distributions $U[\bullet]$ to add uncertainty to our prior choices:

$$p_x \sim \phi_T(0.95, 0.1) \qquad q_x \sim \phi_T(0.5, 0.21)$$
$$p_y \sim \phi_T(0.9, 0.1) \qquad q_y \sim \phi_T(0.7, 0.25)$$
$$\Pr(T) \sim U[0, 1]$$

These assumptions represent that we are reasonably certain that guilty people have no alibis (p_x) and do have a motive (p_y), have a lower and less-certain belief that innocent people have no alibi (q_x), believe with considerable uncertainty that many innocent people will have a motive (q_y), and have a highly uncertain *a priori* belief that the suspect is guilty $\Pr(T)$.

The posterior beliefs about a suspect's guilt implied by this evidence are illustrated by Figure 5 and saved at https://tinyurl.com/bdhytzst. We produced this belief distri-

Figure 5: Posterior inference for suspect with motive and opportunity



bution by drawing one hundred thousand samples from each of the distributions for p_x, p_y, q_x, q_y, and $\Pr(T)$, then calculating $\Pr(T|x = x_0, y = y_1)$ for each one. Our median belief is that 74 out of every 100 suspects with a motive and no alibi will be guilty; that represents a reasonably high assessment. And yet, our confidence in this assessment is relatively low. As we see in the boxplot, we also believe that there is a 25% chance that fewer than half of suspects are guilty under these conditions. If avoiding "false positive" results (mistaken conclusions of guilt) is of paramount importance, we would not find this evidence persuasive.

Now, suppose we are somewhat more certain of our interpretation of the evidence and our prior beliefs:

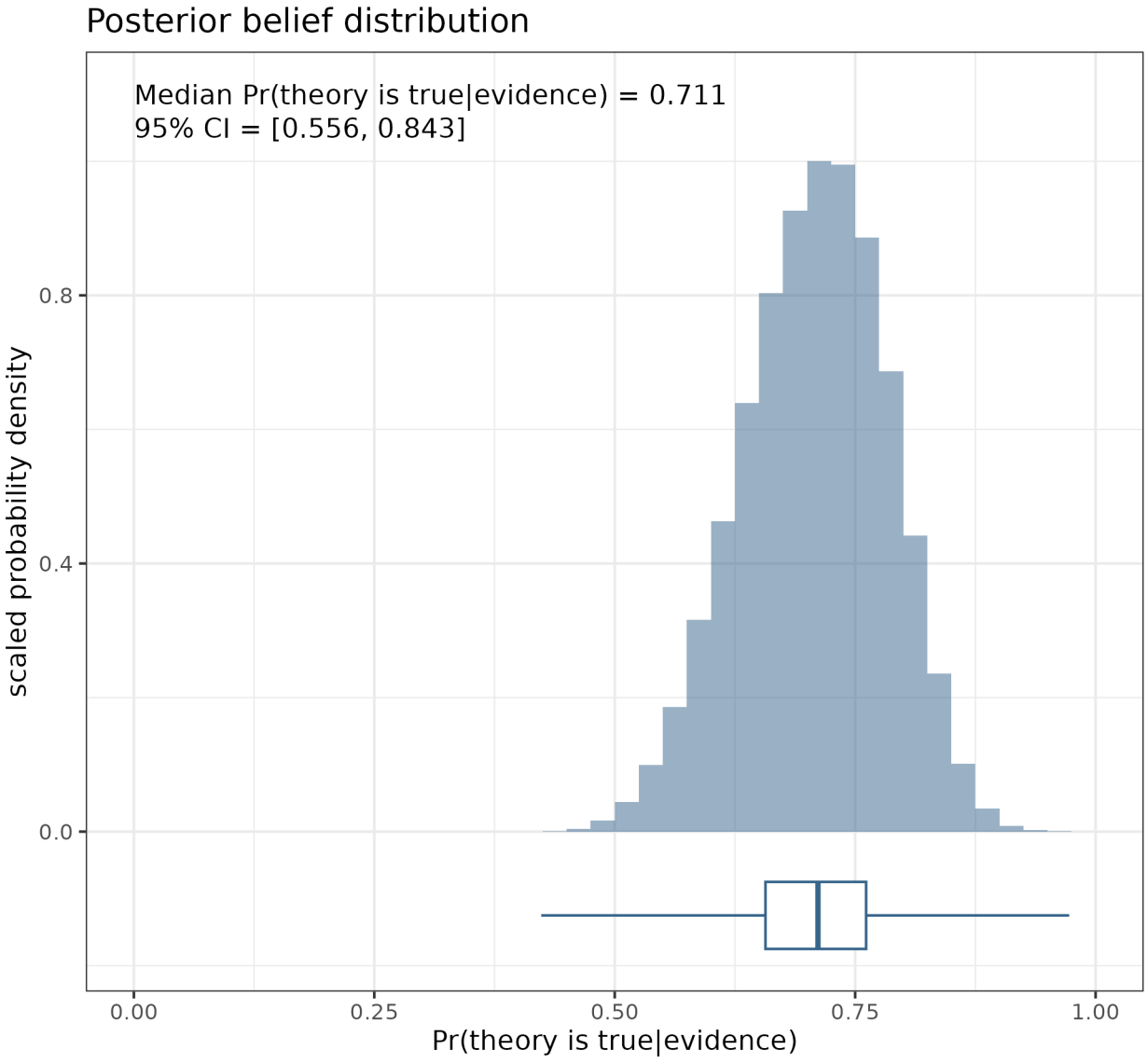
$$p_x \sim \phi_T(0.95, 0.05) \qquad q_x \sim \phi_T(0.5, 0.11)$$
$$p_y \sim \phi_T(0.9, 0.05) \qquad q_y \sim \phi_T(0.7, 0.11)$$
$$\Pr(T) \sim U[0.4, 0.6]$$

The expected values of all five parameters are the same as before. However, we have increased our confidence in all five. For example, we previously stated that our prior belief about the suspect's guilt was $\Pr(T) \sim U[0, 1]$, so that $E[\Pr(T)] = 0.5$. This is still true in our revised assumptions, but now $\Pr(T) \sim U[0.4, 0.6]$; we are much more certain that the suspect is about 50% likely to be guilty before we see any evidence.

These revised assumptions translate into the same expected belief as before, albeit a slightly larger median belief. But our certainty in this belief is much stronger, as shown by Figure 6. Now, we think there is only a very small chance that fewer than half of suspects are guilty under these conditions.

What is the practical difference in the beliefs represented by Figure 5 and Figure 6? Both say there is the same chance that a theory is true—that we are "right," if our goal is to provide evidence for that theory. But Figure 5 illustrates that there is a chance that we are *very* wrong in a way that isn't as possible for the beliefs in Figure 6.

Figure 6: A more confident belief in the same probability of guilt



One way of conceptualizing the difference between the two distributions is to consider the decision efficiency loss created by using less certain beliefs. If one could choose to have the beliefs represented by Figure 5 or Figure 6, assuming both would be equally accurate when we chose them, which would be preferable? Let our posterior belief about $\Pr(T|x)$ be denoted t and its true value be equal to $t^\star$. A typical approach (French and Insua, 2000, Chapter 6) is to presume that decision makers experience losses $L(t)$ by using inaccurate beliefs, where $L(t^\star) = 0$ (that is, we experience no loss when our beliefs are accurate). The nature of the loss depends on the nature of the specific decision being made, but a common assumption is that losses are quadratic in the distance between the correct and erroneous belief:

$$L(t, t^\star) = (t - t^\star)^2 \tag{5}$$

When the $t^\star$ is stochastic so that $t^\star \sim h(t)$, then for any single event a true value of $t^\star$ is drawn from $h(t)$ and then the outcome is chosen from $\{T, \neg T\}$ according to $\Pr(T) = t^\star$. We must make a decision before the particular realization of $t^\star$ is known, and so the loss depends on our choice as well as the particular realization of $t^\star$, $L(t, t^\star)$. The expected loss of a choice t is:

$$\int L(t, t^\star) h(t^\star) dt$$

When the loss is quadratic (as in equation 5), the loss-minimizing decision is $E[t] = \int t * h(t) = \tau^\star$ (see Propositions 20 and 23 on pp. 149-152 of French and Insua, 2000). Thus, the expected loss is:

$$\begin{aligned} E[L] &= \int (E[t] - t^\star)^2 * h(t^\star) dt^\star \\ &= \int (t^\star - E[t])^2 * h(t^\star) dt^\star \\ &= \mathrm{Var}[t^\star] \end{aligned}$$

Consequently, more certain beliefs (i.e., lower variance beliefs) about t are preferable in

that they minimize expected decision losses. We therefore have good reason to perceive the beliefs of Figure 6 as being different than those of Figure 5, and to prefer the more confident beliefs from Figure 6.

Adaptability to multiple forms of Bayesian Process Tracing: alternatives and logical exhaustiveness

As alluded to in the introduction, there are methodological disagreements about the optimum way to perform Bayesian Process Tracing. Our software is designed to accommodate these disagreements by providing accurate inferences under multiple frameworks. For example, Fairfield and Charman (2022, p. 81) stress the importance of comparing a theory to a specific alternative that is mutually exclusive and logically exhaustive. On the matter of specificity, they say:

> ...we must compare two or more well-specified competing hypotheses. Scientific inquiry always involves provisional inference to the best existing explanation. Accordingly, there is little sense in rejecting a hypothesis unless we have an alternative to consider. ...The proposition $\overline{H}$ = *Hypothesis H is not correct* may be too broad or too vague a conjecture on which to condition our reasoning. In other words, it may be extremely difficult or even impossible to assess what kinds of evidentiary predictions $\overline{H}$ might make and with what likelihoods.

This rules out testing a theory against its logical complement, sometimes associated with the "null hypothesis," that the theory is false (and ergo its predictions are false). This standpoint conflicts with the venerable and influential framework of falsificationist epistemology and the associated null hypothesis significance testing framework which focuses on determining whether the testable predictions of a theory are consistent with observed evidence, rejecting those theories whose predictions are falsified. From this

standpoint, which has informed empirical practice in political science for decades (King, Keohane and Verba, 2021), one need not compare a theory to another theory but need only determine whether a theory's predictions are supported or not. But not all objections to specifying concrete alternatives are falsificationist. As one example, Humphreys and Jacobs (2023, pp. 156 and 361; see also Humphreys and Jacobs, 2015) argue for fitting parameters (specifically, counterfactual causal relationship types) to a single directed acyclic graph model when conducting "confirmatory process tracing" without explicit comparison to alternative models.

Although it is unlikely that a researcher has truly considered all alternatives, Fairfield and Charman (2022, p. 84) also believe that as a practical matter a researcher must act as though the theories they compare as logically exhaustive:

> ...we must explicitly specify the leading alternatives we wish to compare at the present stage of research. We then *tentatively* treat this set of hypotheses as exhaustive while we apply the Bayesian inferential framework. If and when new hypotheses emerge, we are always free to add them to the list...

They believe this assumption is benign because they argue for working with log odds ratios comparing the relative probability of two theories, which "makes the exclusivity assumption essentially irrelevant in practice" (p. 85). The log odds ratio transforms Bayesian posterior beliefs for the truth status of two theories, $\Pr(T_1|x)$ and $\Pr(T_2|x)$, into the equivalent ratio:

$$\frac{\Pr(T_1|x)}{\Pr(T_2|x)} = \frac{\Pr(x|T_1)\Pr(T_1)}{\Pr(x|T_2)\Pr(T_2)}$$
$$\log\left[\frac{\Pr(T_1|x)}{\Pr(T_2|x)}\right] = \log\left[\frac{\Pr(x|T_1)}{\Pr(x|T_2)}\right] + \log\left[\frac{\Pr(T_1)}{\Pr(T_2)}\right]$$

This logged Bayes factor allows us to cancel out the typical normalizing denominator of $\Pr(x)$, the total probability of observing evidence x under any state of the world, and eliminating the explicit need to state an exhaustive denominator. If two theories are mutually

exclusive—that is, if only one theory from the pair can be true—then log-odds Bayes' factors can be used to examine the degree to which evidence supports one theory over the other (Fairfield and Charman, 2017). Larger numbers tend to favor T_1 over T_2. Fairfield and Charman (2017, p. 372) recommend interpreting the logarithm of this ratio in a way analogous to the decibel scale for loudness of sound, with $\log_{10}(BF) \geq 30$ (metaphorically, when the evidence speaks louder than 30 decibels) indicating that T_1 is "strongly favored" over T_2. This technique cannot be employed if T_1 and T_2 are not mutually exclusive.

Without taking a position on the optimum way of conducting BPT, our software accommodates multiple approaches. A user can compare mutually exclusive theories $\{T_1, T_2\}$ using log-odds Bayes Factors, as recommended by (Fairfield and Charman, 2022), but they can also specify prior probabilities (as a Dirichlet distribution on a simplex) for $\Pr(T_1, \neg T_2)$, $\Pr(\neg T_1, T_2)$, and $\Pr(\neg T_1, \neg T_2)$ under the restriction that these quantities sum to one. The user taking this approach must then specify for each piece of evidence x distributions for $\Pr(x|T_1, \neg T_2)$, $\Pr(x|\neg T_1, T_2)$, and $\Pr(x|\neg T_1, \neg T_2)$.

Users can also compare a single theory T_1 to its logical complement $\neg T_1$, a form of falsificationist approach to inference, if they so desire. As Fairfield and Charman (2022, p. 81) point out, this necessitates choosing a distribution for $\Pr(x|\neg T_1)$ which can be challenging given its vague nature. However, treating $\Pr(x|\neg T_1)$ as a distribution rather than a point probability makes this process more realistic. For example, if we are completely agnostic about the consistency of evidence with an indeterminate null, we can set $\Pr(x|\neg T_1) \sim U[0, 1]$, a flat uniform distribution giving equal weight to all possible values.

Mutual exclusivity and partially rival hypotheses

Perhaps more controversially, Fairfield and Charman (2022, pp. 86) stress that theories must be mutually exclusive:

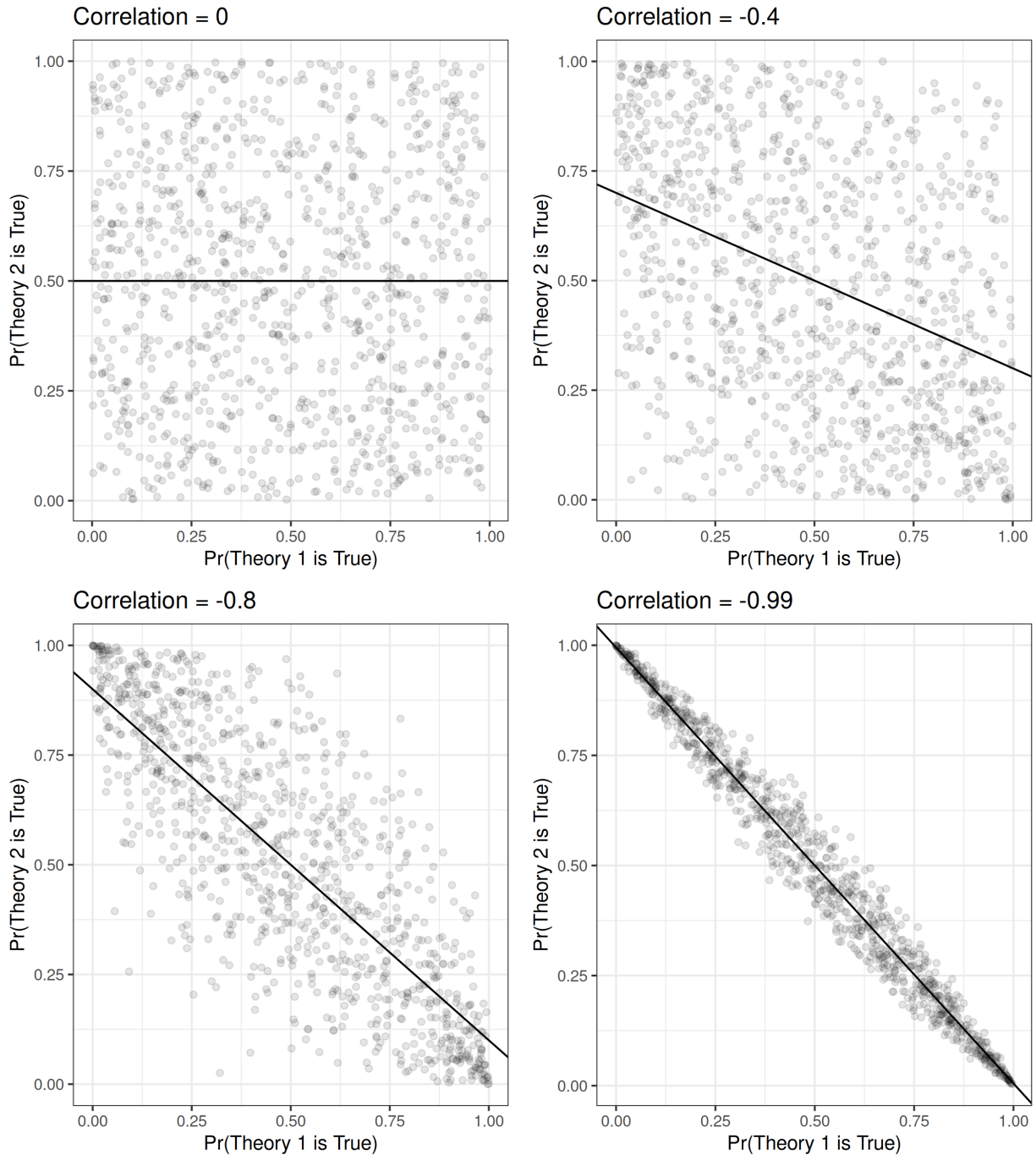
> We stress as a matter of fundamental importance that is is neither sensible

> nor feasible to conduct inference unless we ensure that our hypotheses are mutually exclusive, meaning that only one can be true. If we are not comparing mutually exclusive hypotheses—which we take to be synonymous with rival hypotheses, as is standard throughout essentially all scientific and statistical practice—then we are not testing explanations that can meaningfully be treated as alternative possibilities, and it would hardly make sense to ask which one provides the best explanation.

As they correctly note (p. 87), "social scientists are often uncomfortable with the suggestion that hypotheses should be mutually exclusive" because this may "preclude the possibility that multiple causal factors could simultaneously contribute to the outcome or phenomenon"; this is a key criticism of Zaks (2021, p. 67). Fairfield and Charman respond by noting that, if two causal pathways are potentially operating simultaneously on the dependent variable in a single case, a researcher can wrap these multiple pathways into a single theory and compare it to a theory specifying only one pathway or some other set of pathways entirely.

Continuing our running example of a criminal investigation, it will often be the case that one suspect's guilt tends to exculpate other suspects because we have little *a priori* reason to suspect a multi-person conspiracy. Thus, as our confidence in one suspect's guilt rises, our confidence in other suspects' guilt should fall. However, even completely certain knowledge that one suspect is guilty usually does not definitively rule out the guilt of other suspects, particularly if some evidence tends to inculpate them in the crime. Thus, the theory that one particular suspect is guilty is *partially rival* with the theory that other suspects are guilty. Symmetrically, there are also *partially complementary theories* where the truth of one will tend to be positively associated with the truth of the other. This possibility is not so easily handled by Fairfield and Charman's suggestion of composite theory testing, as we wish to allow the same evidence to simultaneously update our beliefs about multiple suspects' guilt at the same time.

Figure 7: Partially rival theories



Consider the situations shown in Figure 7. Here, we show samples from four different probability densities for the truth status of two theories that we are assessing with the same evidence. However, these theories are only *partially* rival; their truth status is interdependent. Negative correlations signal partially rival hypotheses, with each theory's truth status being negatively associated with the probability of the other's; positive correlations $\rho > 0$ signify partially complementary theories, where greater confidence in the truth of one theory increases our belief in the truth of the other. Stronger correlations indicate a stronger interdependence between the truth status of the theories.

We begin with two partially rival or complementary theories, $\{T_1, T_2\}$. Let $\tau_i = \Pr(T_i|x)$. To sample from the joint density $f(\tau_1, \tau_2|x)$, we alternate sampling from $f(\tau_1|\tau_2, x)$ and $f(\tau_2|\tau_1, x)$ to build up samples from this joint density. We do this by constructing Markov chains of length K, with the kth entry (τ_{1k}, τ_{2k}). After setting initial values for τ_{11} and τ_{21}, we draw the following values sequentially from the appropriate conditional distribution:

$$\tau_{1k} \sim f\left(\tau_1|\tau_{2(k-1)}, x\right)$$
$$\tau_{2k} \sim f\left(\tau_2|\tau_{1k}, x\right)$$

The method, called Gibbs sampling, is a familiar tool from Bayesian statistics. The value of τ_j upon which the distribution of τ_i depends is fixed according to the previous values of the chain. As long as the chain is ergodic, it will converge to the appropriate limiting distribution as $K \to \infty$ (Robert and Casella, 2004, pp. 343-353).

Ergo, we aim to sample from the distribution of $\Pr(T_i|x, \tau_j)$ for $i \in \{1, 2\}$. This expression can be rewritten with Bayes' Rule as:

$$\Pr(T_i|\tau_j, x) = \frac{f(\tau_j, x|T_i)\Pr(T_i)}{f(\tau_j, x|T_i)\Pr(T_i) + f(\tau_j, x|\neg T_i)(1 - \Pr(T_i))} \tag{6}$$

We then use a probability identity to rewrite $f(\tau_j, x|T_i)$ as $f(\tau_j|T_i, x)\Pr(T_i|x)$, giving us:

$$\Pr(T_i|\tau_j, x) = \frac{f(\tau_j|T_i, x)\Pr(x|T_i)\Pr(T_i)}{f(\tau_j|T_i, x)\Pr(x|T_i)\Pr(T_i) + f(\tau_j|\neg T_i, x)\Pr(x|\neg T_i)(1-\Pr(T_i))} \tag{7}$$

We already know how to simulate values of $\Pr(x|T_i)$, $\Pr(x|\neg T_i)$, and $\Pr(T_i)$. What remains is to calculate the value of $f(\tau_j|T_i, x)$ given the current value of τ_j in the Markov chain; call this τ_j^0. Recalling that both τ_i and τ_j are already conditional on x, we can use Bayes' Rule again to write:

$$f(\tau_j^0|T_i, x) \propto \Pr(T_i|\tau_j^0, x) f(\tau_j^0|x) \tag{8}$$

Note that the first term on the right hand side is the left side of Equation 6 again, but will be calculated using information from the other parameter in the Markov chain. Because the marginal distributions are uniform, $f(\tau_j|x)$ is a constant. If τ_i and τ_j are related through a normal copula function with uniform marginal densities so that the conditional expectation $\Pr(T_i|\tau_j^0, x) = 0.5 + \rho(\tau_j^0 - 0.5)$, with ρ being the correlation between τ_i and τ_j as depicted in Figure 7, then we can write:

$$\begin{aligned}\Pr(T_i|\tau_j^0, x) &= \int_0^1 f(\tau_i|\tau_j^0, x)\tau_i \, d\tau_i \\ &= 0.5 + \rho(\tau_j^0 - 0.5)\end{aligned} \tag{9}$$

By substituting Equation 9 into the appropriate term[2] in Equation 8 and then substituting Equation 8 into the corresponding term in Equation 7, we get:

$$\begin{aligned}\Pr(T_i|x, \tau_j) = \frac{(0.5 + \rho(\tau_j^0 - 0.5))\Pr(x|T_i)\Pr(T_i)}{(0.5 + \rho(\tau_j^0 - 0.5))\Pr(x|T_i)\Pr(T_i) + \\ (0.5 - \rho(\tau_j^0 - 0.5))\Pr(x|\neg T_i)(1-\Pr(T_i))}\end{aligned} \tag{10}$$

[2]Symmetrically, we also substitute $\Pr(\neg T_i|\tau_j^0, x) = 0.5 + \rho((1-\tau_j^0) - 0.5) = 0.5 - \rho(\tau_j^0 - 0.5)$ in for $f(\tau_j|\neg T_i, x)$.

Noting that the constants in Equation 8, including the neglected denominator, will divide out.

Thus, we take a single draw from $f(\tau_1|x, \tau_2)$ by simulating a draw from each component of Equation 10 with $i = 1, j = 2$ and setting τ_2 equal to its previous draw in the Markov chain. Then, we take a single draw from $f(\tau_2|x, \tau_1)$ by simulating another draw from components of equation 10 with $i = 2, j = 1$ and setting τ_1 equal to the value we just drew. By repeating this process many times, we build up a sample of $f(\tau_1, \tau_2)$.

An example inference for partially rival and partially complementary theories

Continuing our running example, suppose there are two suspects who may have committed a crime. T_1 is that the first suspect is guilty; T_2 is that the second suspect is guilty. We further believe that the suspects' guilt is interdependent. In this example, we consider two pieces of evidence, x and y. These pieces of evidence tend to incriminate suspect 1 and weakly exonerate suspect 2. For example, suppose that x indicates that suspect 1 has no alibi; this is highly consistent with the first suspect being guilty and relatively inconsistent with the second suspect being guilty, although there is some possibility that the two suspects are collaborating in some way. Evidence y indicates that suspect 1 has a motive to commit the crime. Finally, we think that suspect 2 may be involved with the crime, but our prior belief about this involvement is quite uncertain. We represent the situation

formally as:

$$p_{1x} \sim \phi_T(0.95, 0.05) \qquad q_{1x} \sim \phi_T(0.5, 0.1)$$
$$p_{1y} \sim \phi_T(0.9, 0.05) \qquad q_{1y} \sim \phi_T(0.7, 0.1)$$
$$\Pr(T_1) \sim U[0.4, 0.6]$$

$$p_{2x} \sim \phi_T(0.35, 0.1) \qquad q_{2x} \sim \phi_T(0.7, 0.1)$$
$$p_{2y} \sim \phi_T(0.4, 0.1) \qquad q_{2y} \sim \phi_T(0.8, 0.1)$$
$$\Pr(T_2) \sim U[0.3, 0.7]$$

Above, $p_{ix} = \Pr(x|T_i)$ indicates the conditional probability that evidence x occurs when T_i is true while $q_{ix} = \Pr(x|\neg T_i)$ is that same probability when T_i is false.

We inputted the evidence from this example, with suspect 1 having no alibi and a motive to commit a crime, saving the results at https://tinyurl.com/444j4hzb. These inputs produced the posterior belief distribution shown in Figure 8 using the app. For this computation, we set $\rho = 0.8$. The left panel shows the posterior belief density for the theory that Suspect 1 is guilty, while the right panel shows the posterior belief density for the theory that Suspect 2 is guilty.

Figure 8: Posterior inference for partially rival theories, ho = 0.8)

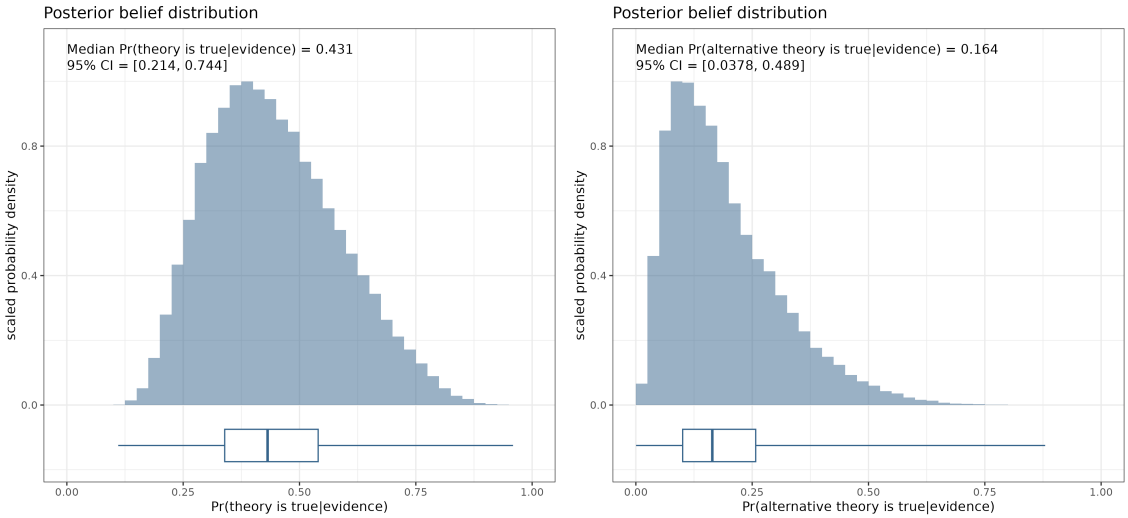
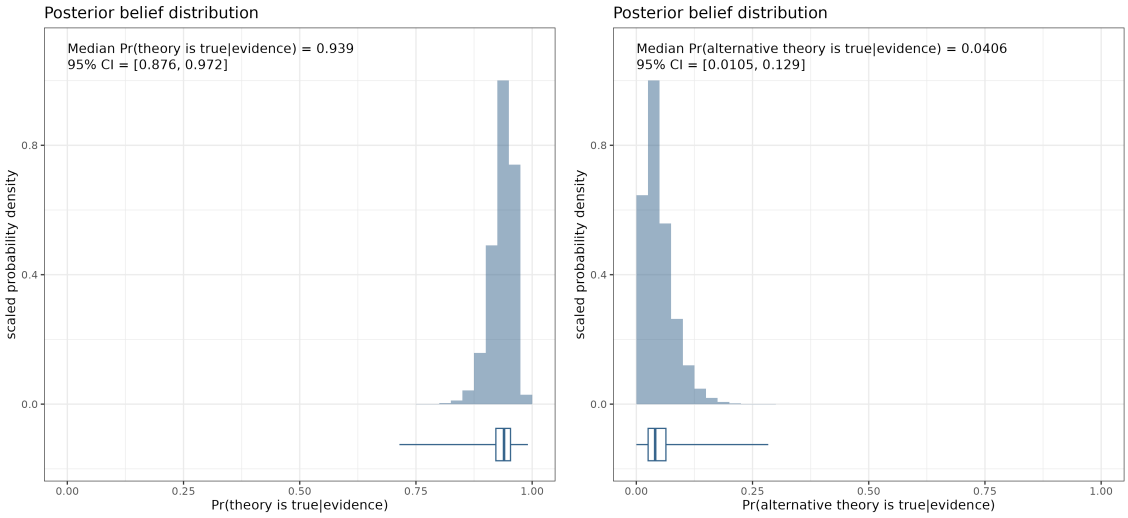


Figure 8 shows the result of an analysis of this evidence presuming different values of correlation between the guilt of the two suspects. If we think that the theories are partially complementary—that is, the guilt of suspect #1 is positively associated with the guilt of suspect #2—then the evidence weakly exonerates both suspects precisely because it is only weakly supportive of suspect #2's guilt, and because we have reason to believe that it is unlikely that only one suspect is guilty. However, if we know that the theories are partially rival ($\rho = -0.8$), the evidence much more strongly incriminates suspect #1 and exonerates suspect #2; this is shown in Figure 9, a version of Figure 8 with ρ set to -0.8.

Figure 9: Posterior Inference for partially complementary theories, ho = -0.8)



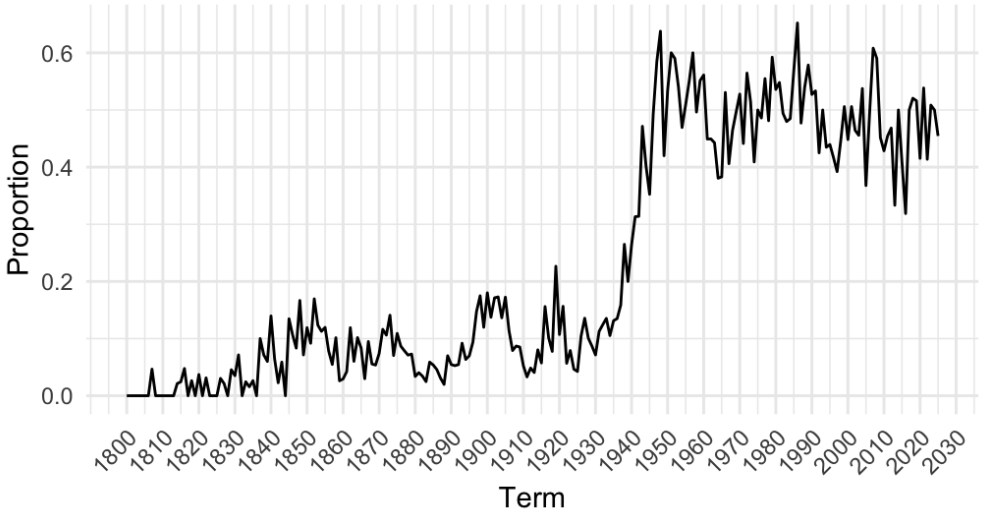
Application

We can use this approach to revisit a famous institutional puzzle studied by scholars of the U.S. Supreme Court. The current Court operates in a hyperpolarized political environment and demonstrates a high degree of dissensus. For example, in the 2025 October Term (October 1, 2025, through June 30, 2026), the Court published 66 merits decisions, 30 of which (45.5%) included at least one dissenting opinion. This level of dissensus is not unusual for the modern Court. In the last 40 years, the proportion of decisions with

dissent has ranged from 31.9% (October Term 2016) to 65.22% (October Term 1986).[3]

But dissensus has not always been the norm. For the first 140 years of the Court's history, the Court's decision-making was governed by a norm of consensus (Walker, Epstein and Dixon, 1988). Before then, dissent was rare (see Figure 10). In fact, prior to 1930, the average dissent rate per term was 6.7%, ranging from a minimum of 0% in the early 1800s to a maximum of 22.7% in 1919. In fact, during Chief Justice John Marshall's tenure, only 83 of the 1,176 cases decided (7.1%) included dissent.

Figure 10: Proportion of Decisions Accompanied by Dissent (1800-2025)



The empirical trend is clear: At some point during the interwar period and the ensuing Great Depression (1917-1941), the Court's opinion-writing norms shifted dramatically, producing the now-common practice of writing separate opinions and openly dissenting. But why? Generations of scholars have sought to explain the causal mechanisms leading to this radical institutional shift. This enduring question (and lack of a satisfyingly clear scholarly answer, as discussed below) makes this question particularly well-suited to examination through Bayesian Process Tracing.

In the following section, we provide a context for our application and identify three competing explanations examined in the literature as potential causal mechanisms: (1)

[3]Data about Supreme Court was accessed from (Spaeth et al., 2026).

membership change resulting in increased ideological distance between the judges, less judicial experience, and the appointment of more dissent-prone justices; (2) increased docket control made possible by the 1925 Judges' Bill; and (3) Chief Justice Stone's leadership style.

After providing context, we use Bayesian Process Tracing, as implemented through the software discussed in the previous sections, to evaluate these competing theories. In particular, we will examine not only how these three theories stack up against each other but also how we can use this approach to explore the interrelatedness of these theories transparently and with the ability to be replicated.

Setting the Stage: Background on the New Deal Court and the Rise of Dissent

By most accounts, modern dissent emerged from the crucible of the Supreme Court's conflict with President Roosevelt over the federal government's role in addressing the economic and social effects of the Great Depression. Since the early 1900s, the Supreme Court had increasingly bound the nation in an ever-tightening jurisprudential straitjacket, elevating *laissez-faire* economic theory to both an individual constitutional right—embodied in the notion of "freedom of contract" and substantive due process under the 5th and 14th Amendments (Mason, 1979)—and constricting the federal government's ability to act through a torturously narrow reading of the Commerce Clause. This jurisprudential regime was characterized by the Court's insistence that federal and state governments lacked constitutional or inherent power to regulate economic dealings between individuals. During this period, commonly designated as the *Lochner* Era after one of the Court's most famous cases on this topic, the Court struck down both federal and state legislation designed to prevent child labor,[4] and set maximum hours and minimum wages.[5].

The Court's resistance to government involvement in economic affairs became espe-

[4] *See, e.g.*, Hammer v. Dagenhart, 247 U.S. 251 (1918)

[5] *See e.g.*, Lochner v. New York, 198 U.S. 45 (1905); A.L.A. Schechter Poultry Corp. v. United States, 295 U.S. 495 (1935)

cially problematic after 1932, when President Roosevelt and his allies in Congress began enacting wide-ranging economic regulations, collectively referred to as the "New Deal," designed to reorganize the American economy. In the winter of 1935, the Court began striking down major pieces of the New Deal legislation, including the Agricultural Adjustment Act, the National Recovery Act, and the Home Owners' Loan Act of 1933 (Savage, 2011).

The *Lochner* Era persisted until the Court suddenly changed course on cases involving private property and economic regulation in 1937. In two cases, *West Coast Hotel v. Parrish*[6] and *NLRB v. Jones & Laughlin Steel Co.*[7] a 5-4 majority determined that the states and Congress had not exceeded their authority in seeking to set wage and hour laws and regulate commercial activity.

The 1936-1937 term also marked a turning point in *who* was dissenting on the Court. Prior to *West Coast Hotel* and *Jones & Laughlin*, the most common group of dissenters were Benjamin Cardozo, (future Chief Justice) Harlan Fisk Stone, and Louis Brandeis (10 cases), joined occasionally by Chief Justice Hughes (4 cases). The Cardozo/Stone/Brandeis coalition primarily dissented in cases involving taxation, property rights, and the Commerce Clause. These justices were critical of the majority for violating a norm of judicial modesty. That is, the dissenting justices accused the majority of a improperly "inquir[ing] into the motives of a legislative body [and] assum[ing] them to be wrongful."[8] But in *West Coast Hotel* and *Jones & Laughlin*, Justices Van Devanter, Sutherland, McReynolds, and Butler—who were known as the "Four Horsemen"—found themselves in dissent. They argued that arguing that the Court has "depart[ed] from well-established principles" articulated in previous cases that the federal govenrment was incapable of regulating commerce earlier determined by the Court to be "purely local".[9] From this point forward, Justices Cardozo, Stone, and Brandeis were more likely to find themselves in the

[6]300 U.S. 379 (upholding state wage and hour laws for women).
[7]301 U.S. 1 (confirming the constitutionality of the National Labor Relations Act).
[8]United States v. Constantine, 296 U.S. 287 (1935) (Cardozo, J., dissenting).
[9]*Jones & Laughlin*, McReynolds, writing in dissent for himself, Sutherland, VanDevanter, and Butler.

majority.

Two justices who disagreed with the Court's jurisprudential switch—Willis Van Devanter and George Sutherland—retired within a year of these decisions. The final two Horsemen—Pierce Butler and James McReynolds—left the Court in 1939 and 1941.[10] In addition to replacing the Four Horsemen, President Roosevelt made four additional appointments. Former dissenters-turned-majority coalition members Benjamin Cardozo and Louis Brandies departed the Court in 1938 and 1939 respectively.[11] Chief Justice Charles Evans Hughes retired in June 1940 and former dissenter Harlan Fiske Stone was elevated to replace him. Roosevelt additionally appointed James Byrnes to replace McReynolds in 1941, but Byrnes resigned in October 1941 to assist Roosevelt with the war effort.

Roosevelt's replacements tended to be comparatively young "New Dealers" with no prior experience as judges.[12] Most of Roosevelt's appointees were politically experienced, and all were dependable supporters of Roosevelt's vision for the American administrative state. Hugo Black, appointed in 1937, and James Byrnes, appointed in 1941, were sitting U.S. Senators from Alabama and South Carolina, respectively. Frank Murphy, appointed in 1940, was the former governor of Michigan and the U.S. Attorney General at the time of his appointment—as was Robert Jackson. William O. Douglas and Felix Frankfurter were former law-professors-turned-executive-branch-officials. Only Wiley Rutledge, appointed in 1939, had absolutely no political or judicial experience; Rutledge has spent essentially his entire career as a law professor. In short, Roosevelt's nominees were long on political experience but short on the norms developed through experience in the federal judiciary.[13]

[10]Butler died in November 1939 and McReynolds retired in January 1941.

[11]Cardozo died in July 1938 and Brandeis retired shortly thereafter in February 1939.

[12]The only exceptions were Hugo Black, who served as a judge in the Birmingham (Alabama) Police Court from 1910–11 while also maintaining an active private practice, and Frank Murphy, who served as a Judge in the Detroit (Michigan) Recorder's Court while simultaneously teaching as a Professor of Law at the University of Detroit before serving as Detroit's mayor from 1930–1933.

[13]Biographical information on Supreme Court Justices was retreived from the Biographical Directory of Article III Judges (*Biographical Directory of Article III Federal Judges, 1789-present*, 2026).

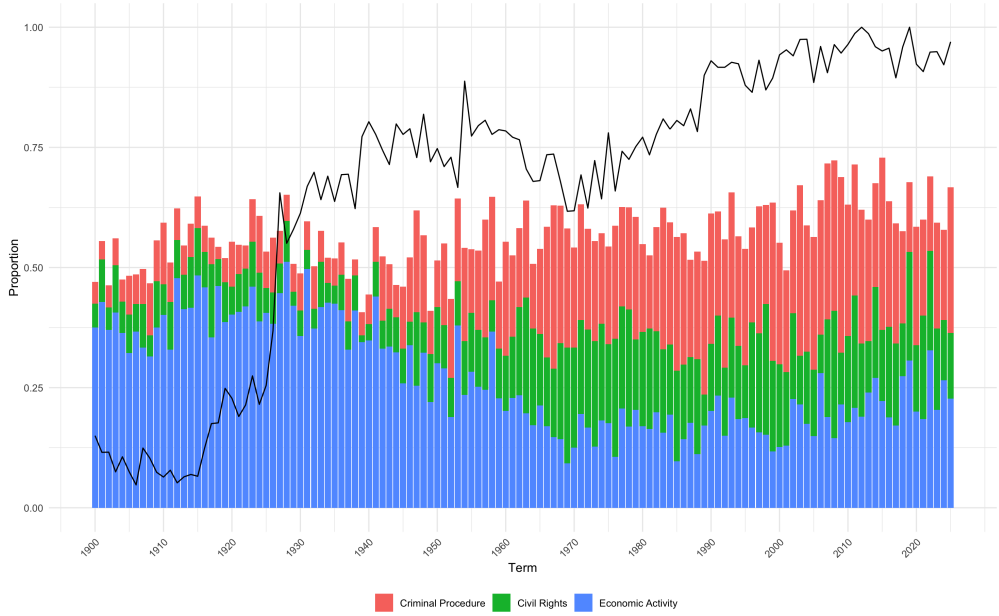
Not only were the new Justices politically savvy and reliable New Dealers, they also were adherents to a growing jurisprudential movement called Legal Realism. Contrary to the previous generation's belief in opinions that relied on the logic of the law, legal realists "believed that in order to understand the law and the behavior of legal institutions, one had to look at individual behavior and use social sciences to find the real causes of particular actions" (Urofsky, 1997, 22). Where the *Lochner* Era court would reason from first principles about the nature of contract between individuals to determine property rights cases, Roosevelt's justices were concerned with the effect of institutions.

The Court's jurisprudential switch and the membership changes occurring in quick succession served to resolve questions over the federal government's ability to regulate the economy. The New Dealers on the Court were then able to shift the content of its docket from economic cases to cases involving individual rights. This transition was made possible only through the creation of an almost entirely discretionary docket in the Judges' Bill of 1925. Figure 11 illustrates this change over time. The black indicates the proportion of Court's docket consisting of discretionary cases. The bars indicate the proportion of the docket consisting of criminal procedure (red), civil rights (green), and economic activity cases (blue). Economic activity cases begin to shrink as a proportion of the docket after 1937, whereas the proportion of the Court's docket comprising criminal procedure and civil rights cases increased dramatically.

Roosevelt's Justices were chosen for their reliable support for New Deal policies. But alignment on economic issues did not require or imply consensus on issues related to civil rights and liberties. The mismatch between commonly used ideology scores and historical evidence highlights this incongruence.

One popular way of exogenously approximating ideology is by leveraging the appointment process to assign DW-Nominate scores. For example, in their study of this question, Hendershot et al (2012) assign to each Justice the average DW-Nominate score of the appointing President's party in the Senate. Figure 12 plots the standard devia-

Figure 11: Change in major areas of the Supreme Court's docket, 1900–2025

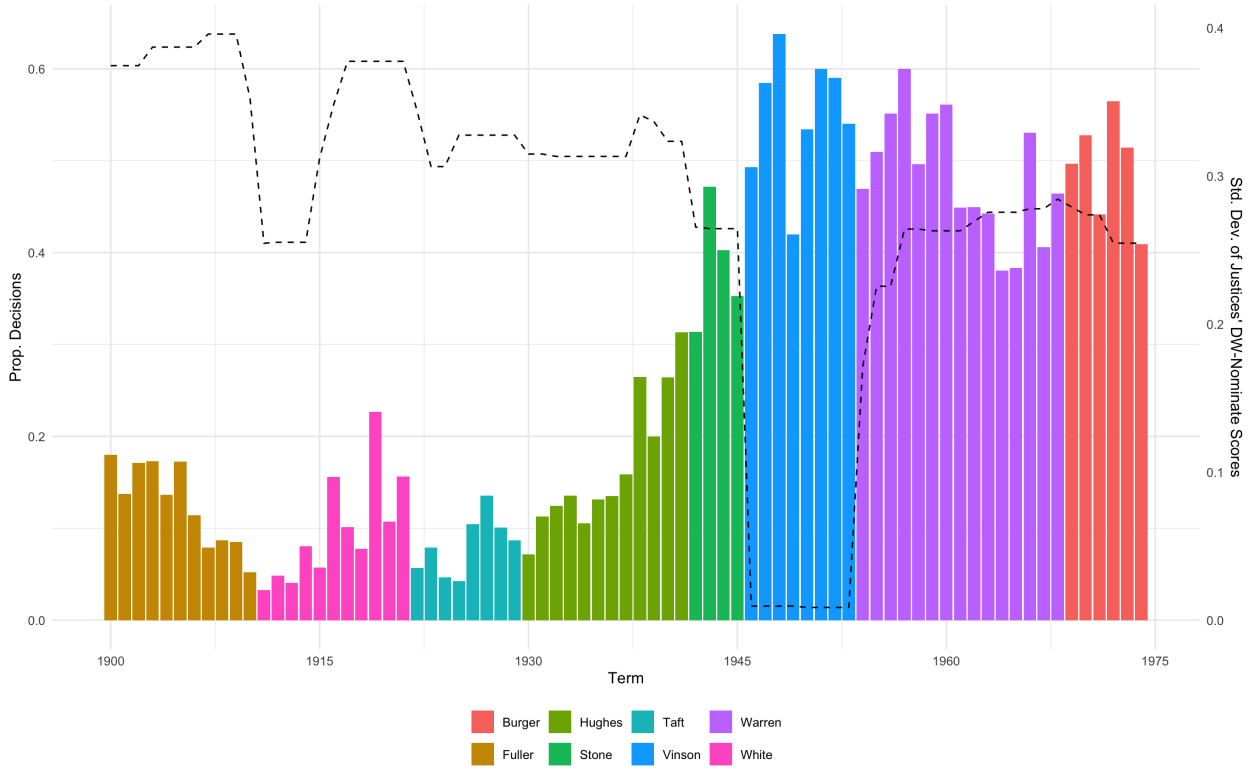


tion of DW-Nominate scores on each natural court (i.e., unique configuration of justices) on the right axis. The bars plotted on the left axis represent the proportion of decisions accompanied by dissent.

As Roosevelt's appointees join the Court, the standard deviation of DW-Nominate scores on the Court shrinks to almost 0—and this ideological convergence occurs at the same time that dissent is hitting historic highs. Why? Evidence from the Justices' papers suggests that, although Roosevelt's appointees were in total agreement on economic issues, they had widely varying ideas about how to interpret constitutional provisions touching on civil rights and liberties (Urofsky, 1997, 30).

The DW-Nominate scores at the time of appointment would not necessarily picked up this ideological element because a focus on civil rights and liberties only came later—after the immediate economic issues spurred by the Great Depression were resolved. Moreover, the New Deal coalition "consisted of three groups that bore the brunt of the Great Depression—white southerners, northern ethnic whites, and the northern white working class" (Abramowitz, 2018, 19). Although unable to exercise much political power, African Americans also began to support Democrats over Republicans (Abramowitz, 2018, 18).

Figure 12: Std. Dev. DW-Nominate Scores by Natural Court Against the Proportion of Decisions Per Term Accompanied by Dissent, 1900-1975



These groups generated a coalition that agreed in a wider role for government intervention. But disagreements over the the appropriate score of civil rights and civil rights and liberties fractured the New Deal coalition. And the Justices' papers suggest that the Court was no different (Urofsky, 1997).

Former Associate—and now Chief—Justice Harlan Fiske Stone led this new Court focused on individual rights. Roosevelt elevated Stone to the Chief Justiceship in 1941 following the retirement of Charles Evans Hughes, a Hoover appointee. Stone, first appointed to the Court as an Associate Justice in 1925, was the only Roosevelt appointee with real judicial experience. Stone was himself a frequent dissenter as an Associate and reportedly chafed at Chief Justice Hughes's method of running the Court's conference discussion. Hughes tended to budget little time for conference and, as a result, "[f]ew of the judges made any extensive notes about the cases they heard; few of them made any careful study of the records or briefs or the cited authorities before they went into conference" (Frank and Mason, 1957, 629, as quoted in Mason (1979), 171). Perhaps in reaction to this, Stone ran a more free-flowing conference that "gave way to endless debates" (Urofsky, 1997, 32). His own attitudes about the propriety and even desirability of dissent transferred over into the manner in which he ran conference (Urofsky, 1997, 31; Walker, Epstein and Dixon, 1988).

In short, between 1937 and 1941, the Court underwent radical ideological, jurisprudential, membership, and leadership changes. After Roosevelt's justices largely captured the Court, they used their discretionary docket to focus heavily on civil rights and liberties while minimizing the number of disputes involving property rights and economic activity. But which of these factors–or to what extent do *all* of these factors explain the dramatic shift in the Court's dissent behavior?

Bayesian process tracing can help us assess which of these possible causal mechanisms best explains the increase in the dissent rate. We start our evaluation with a "hoop test" (Van Evera, 1997, 31; see also Collier, 2011) of the theory that membership change lead to a rise in dissent rates during the late 1930s and early 1940s. In other words, the new Justices' disparate ideological preferences, lack of judicial experience, and individual propensities to dissent produced the permanent change in judicial opinion writing.

This theory is intuitively appealing because it reflects, in part, a simple attitudinal model (Segal and Spaeth, 2002), which forms the foundation of many more sophisticated models of judicial decision-making. This hypothesis echoes the logic that Justices dissent because they had something to disagree about. And not only did they have something to disagree about—their collective experience and beliefs about the nature of law–and their belief in legal realism— informed their approach to judging.

From this starting point, we ask what evidence would it be impossible to see if this theory were true and possible to observe otherwise? If this theory is false, we should not observe ideological homogeneity, a clear expressed desire to adhere to previous judicial norms, and no special independent desire to dissent. If we observe such evidence, then we can conclude that the theory is false.

But what if we do observe this evidence? Such evidence is necessary for the theory to be true, but it is not sufficient. In other words, this evidence may increase the probability that this theory is true, but it is not dispositive.

Here, we would point to evidence from the Justices' papers that Roosevelt's appointees did not, in fact, agree on civil rights issues as evidence of ideological dispersion. We could also point to the Justices' deep political experience and lack of judicial experience as evidence of a lack of adherence to judicial norms.

Our belief about T, that the increase in dissent on the Court in the late 1930s and early 1940s is a product of membership change, now depends on three factors: (1) the

probability that we would observe ideological difference, a lack of judicial experience, and dissent-prone justices during this period given that the emergence of dissent depends on the presence of these factors, $p = \Pr(x = x_1 \mid T)$ (which, in a hoop test, must be equal to 1); (2) the probability that we would observe these elements of membership change on the Court given that the emergence of dissent on the Court was not caused by ideological differences, $q = \Pr(x = x_1 \mid \neg T)$; and our *a priori* belief that the increased dissent was caused by ideological differences among the Justices, $\Pr(T)$. These probabilities would underlie any process-tracing analysis; formalizing them, however, increases transparency and the study's replicability. Presenting the analysis formally also allows for sensitivity testing across variations in assumed probabilities.

Figure 13: Distribution of Potential Posterior Probabilities for all $\Pr(T)$ and $q = \Pr(x = x_1 \mid \neg T)$

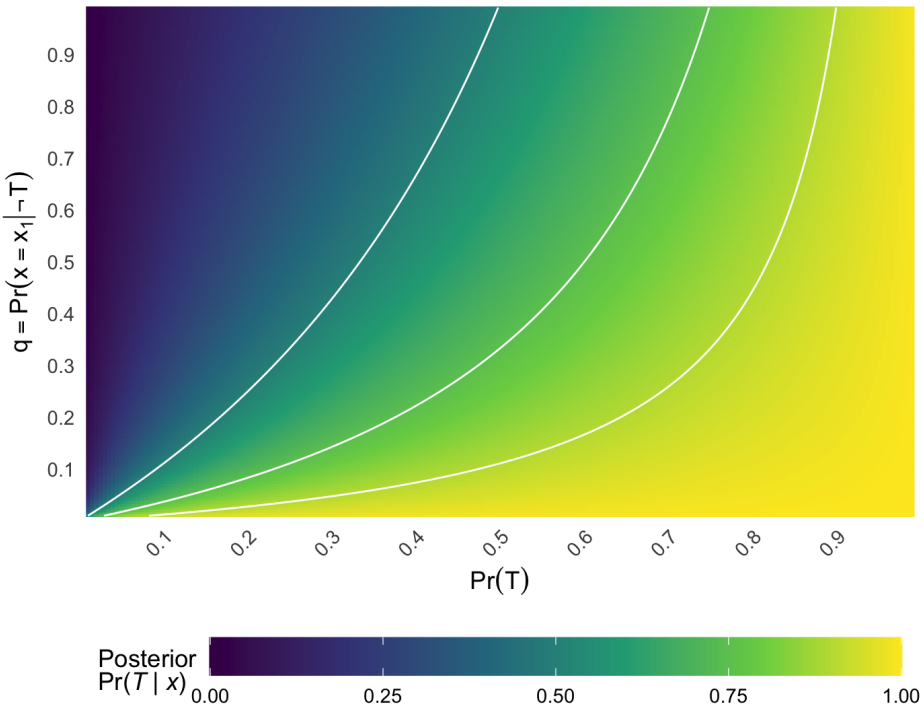


Figure 13 shows how evidence is mapped onto conclusions. The x-axis shows the prior probability $\Pr(T)$. The y-axis shows q, the probability that we would observe all of the characteristics of membership change discussed previously given that the theory was

wrong. This is a hoop test, so $p = 1$. The contour colors in Figure 13 correspond to the posterior belief that our theory is true given the membership-change evidence.

Figure 13 makes clear that the theory that membership change caused permanent and lasting change on the Court is most likely when our *a priori* belief in the theory's truth is high. As our prior decreases, our posterior increasingly depends on the probability of observing ideologically diverse Justices with no judicial experience given that the theory were false.

As appealing and simple as the ideological dispersion explanation is, scholars have offered other compelling explanations for the increase in dissent. In particular, scholars have posited that the Court's ability to control its docket after the passage of the Judges' Bill of 1925 and the ascension of Chief Justice Stone played an important role in fundamentally altering the Justices' opinion-writing behavior.

We can extend our earlier analysis to consider additional evidence. We can define y to be the leadership abilities of Chief Justice Harlan Fiske Stone. As noted above, Chief Justice Stone favored a free-flowing conference, chafed at former Chief Justice Hughes's efforts to marshal the Court to a unanimous opinion, and dissented with relative frequency. Qualitative evidence drawn from the Justices' papers suggests that Stone lacked both the social and task leadership qualities necessary to wrangle what Stone himself called "wild horses" (Urofsky, 1997). This led to long, winding debates and often acrimonious disputes (Walker, Epstein and Dixon, 1988, citing Danelski (1980), Mason (1958), Abraham (1986), Murphy (1964), and Wasby (1984)). One biographer disputes the view that Stone was a poor leader; rather Stone "simply thought that [bringing cohesion to the Court] was not only unnecessary but harmful" (Walker, Epstein and Dixon, 1988, citing Mason (1956)). Stone himself argued that dissent was vitally important to the Court because "[s]ound legal principles . . . never sprang full-fledged from the brains of any man or group of men They are the ultimate resultant of the abrasive force of the clash of competing and sometimes conflicting ideas" (Mason, 1956, 608, 629).

Existing literature on the importance of Justice's policy preferences (e.g., Segal and Spaeth (2002)) suggests that our prior beliefs about the importance of individual policy preferences should be quite high. This would suggest that, perhaps, we expect the probability that our prior belief is correct to be at least greater than 50%. But exiting literature also emphasizes the importance of other factors, such as case salience, as relevant to whether a Justice chooses to dissent. This suggests that our prior should be less than 100%. As a result, any probability greater than 50% but less than 100% is reasonable *given what we knew about judicial behavior before considering the evidence*. As a result, we might say that 70% is a reasonable probability that our prior belief is correct.

Next, we must determine the probability that we would observe the characteristics of membership change (disparate ideologies, lack of judicial experience, and new dissent-prone justices) given that the membership change theory is true, p_x. We then determine q_x, the probability of observing membership change evidence given that the theory is false. If the theory is true, then we are very likely to see these characteristics but not guaranteed, so specifying $p_x = 0.95$ is reasonable.

But, even if the the membership change hypothesis is not true, the political environment at the time makes observing these factors is possible *even if* they did not directly lead to the change in dissent behavior. The nation was at a moment of crisis and partisan realignment. Roosevelt's long tenure, the advanced age the existing Justices, and the political conflict between the President and the Court could have produced the membership change observed but without the additional affect of changing the operation of the Court. Perhaps Roosevelt's justices simply moved the Court to a more reliable ideological position, but the change in the mode (as opposed the content) of the Justices expression is a function of other factors. Still, we would expect to be slightly less likely to see this combination of factors if the hypothesis were not true. As a result, we might conclude that we are just as likely to see the membership change hypothesis as not if the theory were not true. This suggests that $q_x = 0.5$ would be a reasonable probability.

We must next consider our new evidence: what is the probability that we would observe poor leadership by Stone given that the membership change hypothesis is true (p_y) and the probability that we would observe Stone's brand of leadership if the hypothesis is false (q_y)? A leadership style that consciously disdained fostering consensus might be more likely than not under the conditions of membership change. In other words, the wholesale membership change observed on the Court during this time would be more likely to permit such a leadership change than a period of more stable membership. As a result, it is reasonable to conclude that observing leadership such as Stone's during this time is more likely than not and to assign a probability of $p_y = 0.6$.

The probability of observing Stone's leadership style if our hypothesis is false is more challenging to determine. In this situation, the set of plausible alternatives could encompass theories such as that Chief Justice Stone's leadership is responsible for the change in opinion-writing behavior or that the change in dissent behavior is a function of docket change. Any plausible alternative must be considered here when determining q_y.

Figure 14: Distribution of Potential Posterior Probabilities for all $\Pr(T)$ and $q_y = \Pr(y = y_1 \mid \neg T)$

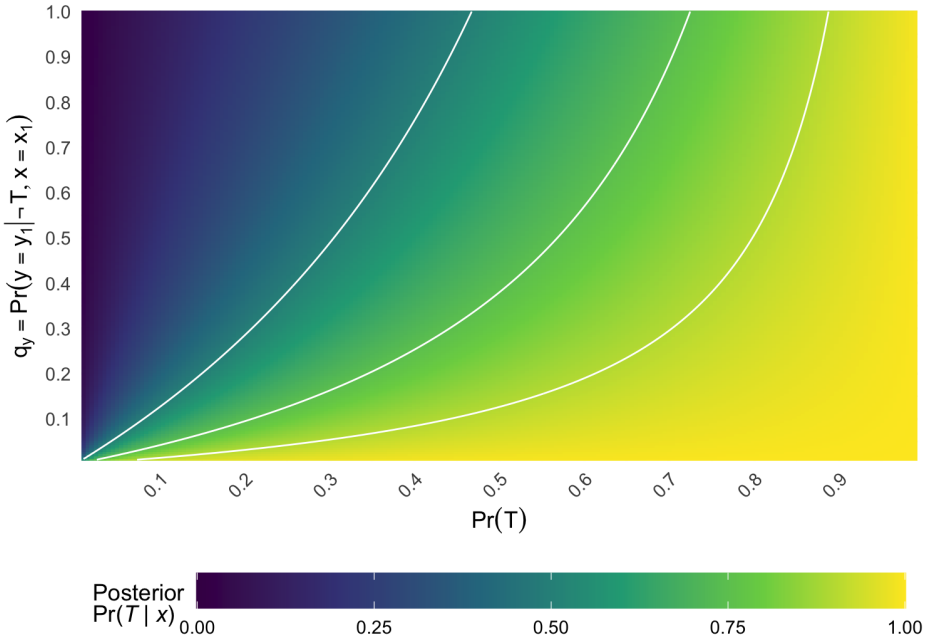


Figure 14 shows the conclusions that we can draw under various assumptions about the *a priori* belief (on the x axis) and the probability of observing Stone's leadership style given that our hypothesis is false (on the y axis), holding all other values in the calculation constant as assumed. If we start with a high prior probability that the hypothesis is true, then almost no value of q_y can dispel us from that notion. When the prior probability $\Pr(T) > 0.9$, the posterior probability $\Pr(T \mid x, y)$ approaches 1, regardless of the value of q_y. But when we start indifferent about the prior probability, q_y can have a large effect on the posterior. For example, if we start with an uninformative prior of 0.5 and assess the probability of observing a leadership style like Stone's given that our theory is not true as relatively high, this reduces our posterior probability to below 0.5. But if we think that it is highly unlikely to observe Stone's leadership style if our theory is true, then the posterior probability approaches 1.

As laid out above, there are five pieces of information that must be supplied and about which we might be uncertain: p_x, q_x, p_y, q_y, and the prior $\Pr(T)$. As probabilities, these are bounded between 0 and 1 and any distributions must be similarly bounded. We therefore use truncated normal distributions $\phi_T(\bullet)$ and uniform distributions $U[\bullet]$ to set this information:

$$p_x \sim \phi_T(0.95, 0.1) \qquad q_x \sim \phi_T(0.5, 0.2)$$
$$p_y \sim \phi_T(0.6, 0.1) \qquad q_y \sim \phi_T(0.7, 0.25)$$

$$\Pr(T) \sim U[0, 1]$$

These assumptions represent that we are reasonably certain we are unlikely to see disparate ideologies on the Court, a lack of judicial experience, and dissent-prone justices if membership change on the Court is the source of the permanent change in the Court's opinion-writing structure (p_x). We have a lower and less-certain belief that we would observe a leadership style like Stone's if membership change was the true cause (q_x), and

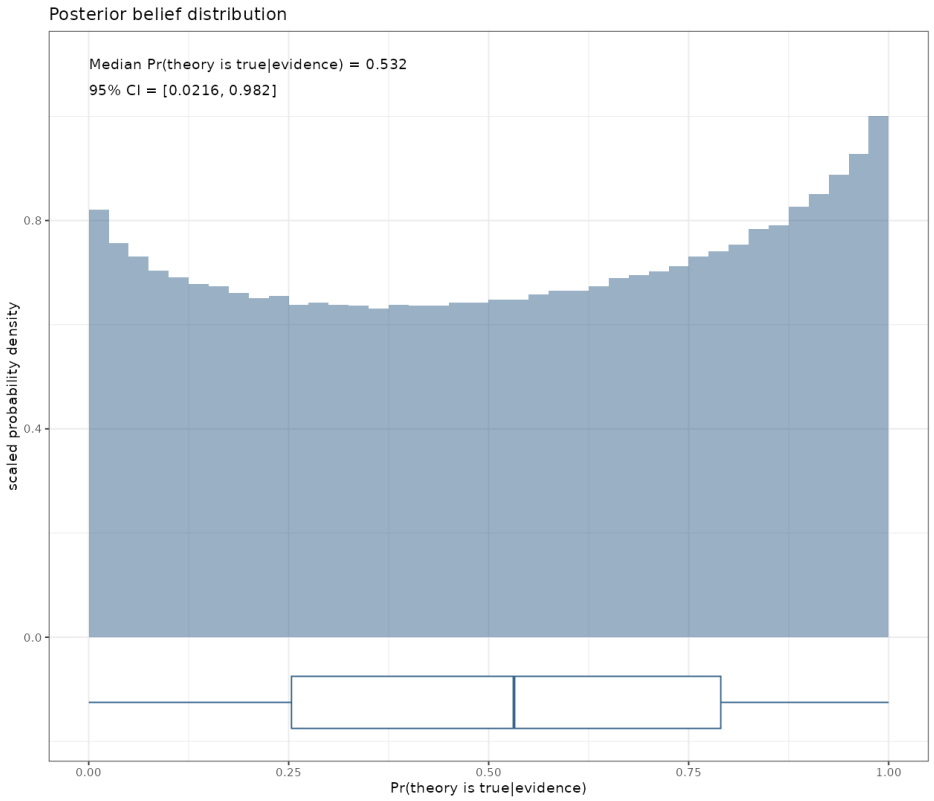
a slightly stronger and more certain belief that we would observe a leadership style like Stone's if membership change is *not* the true cause (because the chief justice's leadership provides a highly plausible alternative hypothesis).

Finally, it is reasonable to have a highly uncertain *a priori* belief membership change is responsible for the changes on the Court ($\Pr(T)$). This is true even if we believe that the underlying probability is relatively high ($\Pr(T) = 0.7$). This is the case for several reasons. First, modern literature on the Court's decision making tends to start in the mid-1940s. This is a function of the available data for many years,[14] as well as the recognition that the Court underwent a dramatic change in the early 1940s. As a result, the bulk of quantitative analyses of the Court reflect the Court in the second half of the 20th century and beyond. It is plausible to think that attitudinal models of the Court were always functioning, even if muted by other institutional factors. But the existing evidence is unclear. As a result, specifying a prior distribution with a wide dispersion (i.e., standard deviation) is reasonable.

The posterior beliefs about the true cause of the Court's institutional change are illustrated by Figure 14. We produced this belief distribution by drawing one million samples from each of the distributions for $p_x, p_y, q_x, q_y,$ and $\Pr(T)$, then calculating $\Pr(T \mid x = x_1, y = y_1)$ for each one. Our median posterior belief is that it is more likely than not that membership change caused the permanent shift in the Court's opinion-writing norms, such that $\Pr(T \mid x, y) = 0.53$). But we lack high confidence about this conclusion. The 95% confidence interval is quite wide, ranging from 0.0216 to 0.982. There is also a 25% that our final posterior probability for the truth of our hypothesis is 0.25. The strength of the additional data provided by biographers' assessment of Stone's leadership—and the Justices' own words in their personal papers—suggest that Stone's leadership was a likely driver of the Court's behavior. Membership change is likely not the *only* important factor.

[14]For example, earlier versions of the Supreme Court database began at the 1946 October Term.

Figure 15: Posterior probability that Membership Change is Responsible for Shift in Opinion-Writing Norms



Conclusion

This paper reviewed how qualitative evidence can be interpreted through the lens of Bayesian inference using Bayesian Process Tracing. To do so, each piece of evidence x must be (qualitatively) evaluated for its consistency with a given theory T. Specifically, the analyst must evaluate the probability of observing x when the theory is true, $\Pr(x|T)$, and the probability of observing x when the theory is false, $\Pr(x|\neg T)$. Along with the prior (pre-evidence) belief that the theory is true $\Pr(T)$, these judgments can be translated via Bayes' rule into a posterior belief about whether the theory is true, $\Pr(T|x)$.

The precision and rationality of this process, and the intrinsic transparency and reproducibility that it provides, are methodologically appealing. But it is more easily said than done in practice: the training required is significant. Nor does extant technique allow for several common complications, such as the simultaneous testing of non-mutually exclusive theories that explain the same event.

The paper offers solutions to both of these problems. First, it introduces a browser-based application that allows anyone to make Bayesian inferences with qualitative evidence. Second, it proposes a copula-based method for allowing interdependence between the truth status of multiple non-exclusive theories. To overcome the analytical difficulties of computing posterior belief densities for statistically dependent theories conditional on multiple pieces of evidence, the paper leverages well-known techniques from Bayesian statistics to build up samples from the target density using Monte Carlo methods.

References

Abraham, Henry J. 1986. *The Judicial Process*. 5th ed. Oxford University Press.

Abramowitz, Alan I. 2018. *The Great Alignment: Race, Party Transformation, and the Rise of Donald Trump*. Yale University Press.

Bennett, Andrew. 2015. Disciplining our conjectures: Systematizing process tracing with Bayesian Analysis. In *Process Tracing: From Metaphor to Analytic Tool*. New York, NY: Cambridge University Press pp. 276–298.

Bennett, Andrew, Tasha Fairfield and David Hillel Soifer. 2019. "Comparative Methods and Process Tracing." American Political Science Association Organized Section for Qualitative and Multi-Method Research, Qualitative Transparency Deliberations, Working Group Final Reports, Report III.1. URL: https://dx.doi.org/10.2139/ssrn.3333405.

Biographical Directory of Article III Federal Judges, 1789-present. 2026.
URL: *https://www.fjc.gov/history/judges*

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Garrick Aden-Buie, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges. 2026. *shiny: Web Application Framework for R*. R package version 1.14.0.
URL: *https://CRAN.R-project.org/package=shiny*

Collier, David. 2011. "Understanding process tracing." *PS: Political Science & Politics* 44(4):823–830.

Danelski, David J. 1980. *The Influence of the Chief Justice in the Decisional Process of the Supreme Court*. John Wiley.

Elman, Colin and Diana Kapiszewski. 2014. "Data access and research transparency in the qualitative tradition." *PS: Political Science & Politics* 47(1):43–47.

Elman, Colin, John Gerring and James Mahoney. 2016. "Case study research: Putting the quant into the qual." *Sociological Methods & Research* 45(3):375–391.

Fairfield, Tasha and Andrew Charman. 2022. *Social Inquiry and Bayesian Inference: Rethinking Qualitative Research*. Cambridge: Cambridge University Press.

Fairfield, Tasha and Andrew E Charman. 2017. "Explicit Bayesian analysis for process tracing: Guidelines, opportunities, and caveats." *Political Analysis* 25(3):363–380.

Frank, John P. and Alpheus Thomas Mason. 1957. "Harlan Fiske Stone: An Estimate." *Stanford Law Review* 9(3):621.
URL: *http://dx.doi.org/10.2307/1226615*

French, Simon and David Rios Insua. 2000. *Statistical Decision Theory*. Wiley & Sons, Limited, John.

Hendershot, Marcus E., Mark S. Hurwitz, Drew Noble Lanier and Richard L. Pacelle. 2012. "Dissensual Decision Making: Revisiting the Demise of Consensual Norms within the U.S. Supreme Court." *Political Research Quarterly* 66(2):467–481.
URL: *http://dx.doi.org/10.1177/1065912912436880*

Humphreys, Macartan and Alan M. Jacobs. 2015. "Mixing methods: A Bayesian approach." *American Political Science Review* 109(4):653–673.

Humphreys, Macartan and Alan M. Jacobs. 2023. *Integrating inferences: causal models for qualitative and mixed-method research*. Cambridge University Press.

King, Gary, Robert O Keohane and Sidney Verba. 2021. *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.

Mason, Alpheus Thomas. 1956. *Harlan Fiske Stone: Pillar of the Law*. Viking Press.

Mason, Alpheus Thomas. 1958. *The Supreme Court from Taft to Warren*. Louisiana State University.

Mason, Alpheus Thomas. 1979. *The Supreme Court from Taft to Burger*. 3rd ed. Louisiana State University Press.

Moravcsik, Andrew. 2014. "Transparency: The revolution in qualitative research." *PS: Political Science & Politics* 47(1):48–53.

Murphy, Walter F. 1964. *Elements of Judicial Strategy*. University of Chicago Press.

Robert, Christian P. and George Casella. 2004. *Monte Carlo Statistical Methods*. Springer Texts in Statistics New York: Springer.

Rohlfing, Ingo. 2012. *Case Studies and Causal Inference: An Integrative Framework*. New York: Palgrave Macmillan.

Savage, David G., ed. 2011. *Guide to the Supreme Court*. 5th ed. CQ Press.

Segal, Jeffrey A. and Harold Spaeth. 2002. *The Supreme Court and the Attitudinal Model Revisited*. Cambridge University Press.

Spaeth, Harold J., Lee Epstein, Michael J. Nelson, Andrew D. Martin, Jeffrey A. Segal and Sara C. Benesh. 2026. "Supreme Court Database.".
URL: *http://scdb.psu.edu*

Urofsky, Melvin I. 1997. *Division and Discord: The Supreme Court under Stone and Vinson, 1941-1953*. University of South Carolina Press.

Van Evera, Stephen. 1997. *Guide to methods for students of political science*. Cornell University Press.

Waldner, David. 2026. *Qualitative Causal Inference and Explanation*. Cambridge: Cambridge University Press.

Walker, Thomas G., Lee Epstein and William J. Dixon. 1988. "On the Mysterious Demise of Consensual Norms in the United States Supreme Court." *The Journal of Politics* 50(2):361–389.
URL: *http://dx.doi.org/10.2307/2131799*

Wasby, Stephen L. 1984. *The Supreme Court in the Federal Judicial System*. 2d ed. Holt, Rinehart, and Winston.

Zaks, Sherry. 2021. "Updating Bayesian (s): A critical evaluation of Bayesian process tracing." *Political Analysis* 29(1):58–74.