Robust and Replicable Bayesian Process Tracing

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

Justin Esarey *Wake Forest University*

Recent scholarship has drawn attention to the utility of Bayesian methods for qualitative research and the need for transparency and reproducibility in application of its methods. It has also pointed out shortcomings, such the high training requirement for using these methods and their inability to compare theories that are not mutually exclusive. This paper introduces free and open source software for robust and replicable Bayesian process tracing. It also introduces a new method for simultaneous assessment of non-exclusive theories using Bayesian updating. The procedure is robust because it allows researchers to test their conclusions for sensitivity to their assumptions and the strength of their evidence. The software facilitates reproducible research because evidence, beliefs, and uncertainty about them are transparently and rigorously mapped into conclusions that can be verified at will. Finally, and perhaps most importantly, the software lowers the barrier to entry of using Bayesian methods by stressing visually intepretable cues.

Keywords: process tracing, Bayesian methods, replicability

Recent scholarship has emphasized the deep connection between qualitative methodology, particularly process tracing, and Bayesian reasoning (Rohlfing, 2012, Chapter 8; Bennett, 2015; Humphreys and Jacobs, 2015). Part of the motivation for this scholarship has been the desire to maximize the transparency, reproducibility, and robustness of research using this method (Elman and Kapiszewski, 2014, Moravcsik (2014), Elman, Gerring and Mahoney (2016)). Explicit, mathematical Bayesian reasoning is important 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).

*__Current version__: November 13, 2024; __Corresponding author__: justin@justinesarey.com.

This scholarship also lays out some barriers to widespread adoption of Bayesian process tracing (BPT). 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). 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).

Another problem is a lack of agreement on how evidence should be translated into conclusions in a standardized way. This point is repeatedly made in the existing literature, particularly as it concerns the conversion of qualitative statements of evidence into quantitative probability distributions. "In 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). Zaks (2021) notes that, "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" (p. 71). She also highlights 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).

Finally, a major outstanding problem in the literature on Bayesian hypothesis testing concerns the assessment of *partially rival* and *partially complementary hypotheses*. These are hypotheses whose truth statuses are neither completely mutually exclusive nor completely unrelated. Current methods of Bayesian process tracing are incapable of simultaneously or comparatively testing these hypotheses (Zaks, 2021, p. 67):

> The Bayesian approach is not equipped to handle the range of forms hypothe-

ses 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.

This paper addresses these problems by developing and presenting new software designed to facilitate full, formal Bayesian process tracing by scholars without extensive mathematical training. The software is freely available on the web, usable by anyone with a browser and an internet connection. It provides a standardized format for specifying hypotheses, entering descriptions of and/or direct links to qualitative or quantitative evidence, and visually specifying the consistency of the evidence with these hypotheses using widely understood graphs. The software can accommodate statistical dependence in interpretation of the evidence (i.e., that the consistency of evidence with a hypothesis depends on other evidence). 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.

The conclusions facilitated by the software are also readily interpretable; it produces graphical presentations of the posterior beliefs about the hypothesis that are consistent with that evidence and with Bayes' rule. Perhaps most importantly, the software allows for easily testing the robustness of conclusions to different intepretations of evidence and possible biases that might have been present in those interpretations. Finally, the software produces a report summarizing the entire process and allowing anyone to reproduce it. The source code is open under a Creative Commons license for those who wish to examine or modify its operation.

The remainder of the paper proceeds as follows. First, it reviews the method of Bayesian process tracing, making note of variations in the literature and laying mark-

ers for what software needs to do. It focuses on how some critiques in the literature can be addressed, especially how new software can lower the start-up cost of using BPT and help researchers ensure their conclusions are robust to different interpretations of the evidence adduced. The paper then presents a new approach to drawing inferences about theories that are not mutually exclusive. Finally, it presents the software and describe how it is used. A simple motivating example common to this literature, the investigation of a crime, is followed through the entire paper.

Applying Bayesian methodology to process tracing

The core logic of process tracing, as described by Van Evera (1997, pp. 64-67), is finding pieces of evidence that are logically consistent (or inconsistent) with casual mechanisms hypothesized by the researcher to be in operation for a case. Bayesian process tracing translates these pieces of evidence into expressions of the probability that the evidence is consistent with that causal mechanism. Finally, it uses Bayes' rule to determine how confident we should be in our hypothesis given that evidence. In this section, we review that process in depth.

As noted in the introduction, not everyone interprets this basic idea in the same way. For example, Humphreys and Jacobs (2015) (p. 653) present a framework for "Bayesian integration of quantitative and qualitative data (BIQQ)" that stresses differentiating types of cases from one another based on their counterfactual response to a treatment (p. 656). The hypotheses being tested in their framework are about parameters linking independent to dependent variables, not about the truth or falsity of overarching theoretical mechanisms. Given these parameters, individual observations have an unknown type that determine whether and how their outcome changes according to their treatment status. Thus, the aim of the BIQQ framework is to produce a belief distribution about a vector of parameters and the unobservable types of cases of interest. This distribution is partially informed by qualitative information about some of the observations used in an otherwise quantita-

tive analysis (p. 660).

By contrast, the procedure laid out by Rohlfing (2012) and Fairfield and Charman (2017)—and the approach we will take in this article—allows for much more mechanism-centered (and qualitative) hypotheses in addition to parametric hypotheses.[1] Consider an example hypothesis about Chilean tax reform offered by Fairfield and Charman (2017, p. 374):

> H_I: *The opposition accepted the reform because Chile's institutionalized party system motivates cross-partisan cooperation and consensual politics*.

In this framework, evidence can also include qualitative statements of findings:

> E_1: *Governing-coalition informants told the investigator that the center-left coalition discussed including a measure to eliminate the tax subsidy in multiple prior tax reforms, but that measure was ruled out as infeasible on every occasion due to resistance from the right coalition*.

Investigators using this framework must specify the degree to which each piece of evidence (including that derived from qualitative case studies) is consistent with the hypothesis, $\Pr(E_1|H_I)$. The same must be done for the consistency of evidence with rival hypotheses (Fairfield and Charman, 2017) or the original hypothesis's negation, the null (Rohlfing, 2012, pp. 189-190). Bayes' rule is then used to translate this evidence and prior beliefs about the hypothesis $\Pr(H_I)$ into updated posterior beliefs about the hypothesis $\Pr(H_I|E_1)$.

Basic Bayesian process tracing

For example, a "hoop test" (Van Evera, 1997, p. 31; see also Collier, 2011) looks for a piece of evidence that can only exist if a theory is false; if that evidence exists, the theory is

[1] See also Fairfield and Charman (2019).

falsified. A canonical example is an alibi, which—if it exists—demonstrates that a suspect could not have committed a crime. However, 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 reasoning is easily 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 binary evidence $x \in \{x_0, x_1\}$, such as in the example of having an alibi or not, then:

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

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:

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

In this case, the "hoop test" is strongly informative under some circumstances. 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:

$$\Pr(T|x = x_1) = \frac{\Pr(x = x_1|T)\Pr(T)}{\Pr(x = x_1|T)\Pr(T) + \Pr(x = x_1|\neg T)(1 - \Pr(T))} = 0$$

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

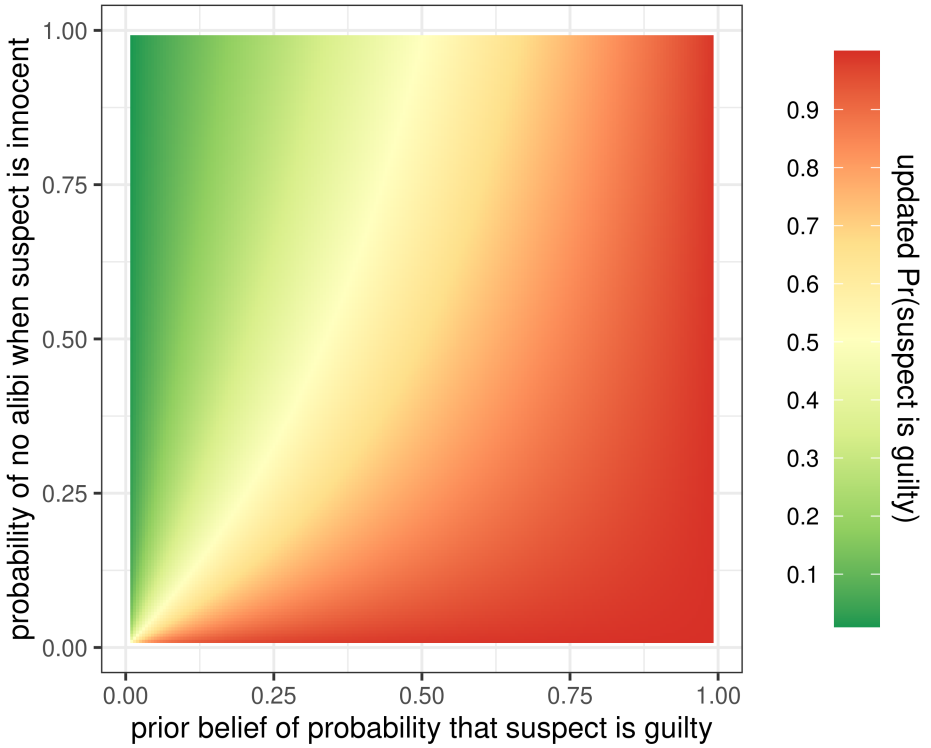
process tracing into Bayesian language:

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

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.

Figure 1 shows how evidence is mapped into conclusions. The x-axis shows the prior probability $\Pr(T)$. The y-axis shows q, the probability of a suspect having no alibi when they are innocent. As this is a hoop test, $p = 1$. The color of the corresponding point on the graph is the updated belief that the suspect is guilty given that they don't have an alibi. The reader can see (and, if desired, reproduce) the analysis connecting the evidence p and q and the prior belief $\Pr(T)$ to the author's ultimate inference about $\Pr(T|x)$ by simply looking at the appropriate plot. Furthermore, if the reader disagrees with the author's interpretation of the evidence (for example, if the reader believes that innocent suspects are more or less likely to have alibis compared to the author), it is easy to assess the sensitivity of the author's conclusion to that interpretation. In this case, the figure makes clear that lacking an alibi is only persuasive evidence of guilt in the implausible scenario where innocent people rarely or never have alibis, or when we already strongly suspected their guilt even before discovering they had no alibi.

Figure 1: Mapping evidence into conclusions (updated posterior beliefs after a passed hoop test)



Compound evidence

The example above considers a single piece of binary evidence with strict, stark implications for the truth of a theory. 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 $y \in \{y_0, y_1\}$:

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

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 1, we need only provide the additional quantities p_y and q_y to determine the updated belief $\Pr(T|x = x_0, y = y_1)$:

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

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

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

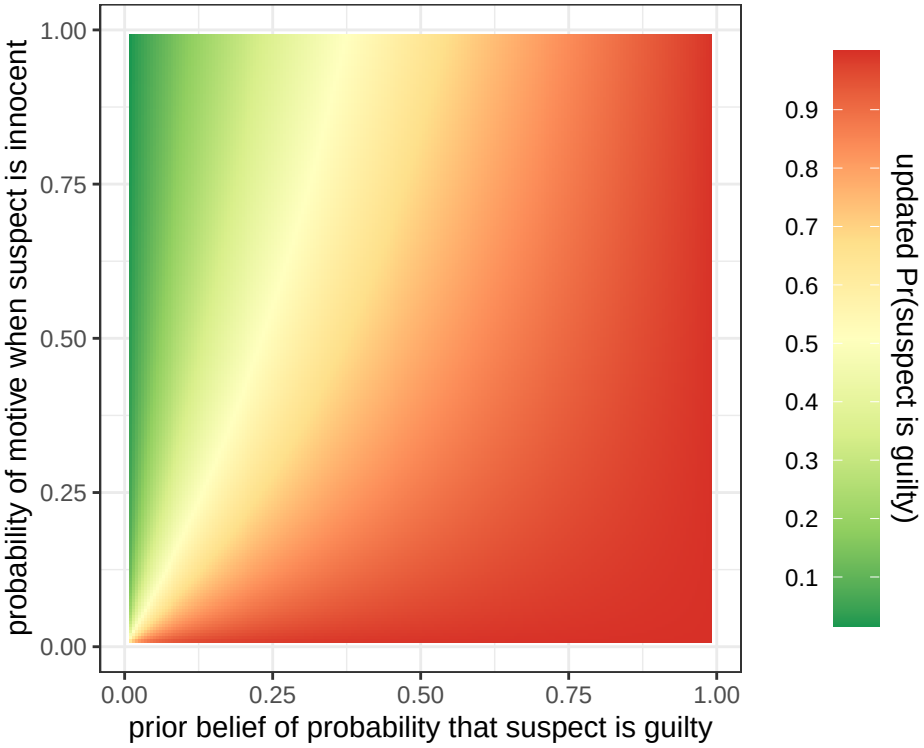
For example, suppose that we believe (based on background information) that there is a 10% chance that a particular suspect committed a crime, that suspect has no alibi, and that suspect also has a motive to commit the crime. To form a belief about their guilt, we must specify:

1. p_x: *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. However, we relax our previous strict assumption of the hoop test by presuming 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 this at 95%.
2. q_x: *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; 50% seems like a reasonable figure.
3. 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%.
4. 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.
5. $\Pr(T)$: *What is our a priori belief that the suspect is guilty?* This answer can also vary widely according to the details of the situation.

Figure 2 shows the conclusion we can draw under various assumptions about *a priori* belief (on the x axis) and the probability of innocent suspects having a motive (on the y axis), setting other values in the calculation constant as assumed. As the figure shows, if

we have a relatively low ($< 12.5\%$) prior belief of guilt, even this combination of evidence is often not particularly incriminating unless almost no innocent suspects have a potential motive.

Figure 2: Mapping compound evidence into conclusions



Uncertainty in interpretation

The above example is limited in several ways. First, although there are five pieces of information that go into the final judgment, it is difficult for us to fully explore the sensitivity of our judgment to all of them simultaneously. Visualizations such as Figure 2 will struggle to show the combined effect of more than two or three changes at once. Second, although Figure 2 can show the relationship between p_y, the prior $\Pr(T)$, and the final conclusion about $\Pr(T|x, y)$, that conclusion still assumes fixed values for p_y and $\Pr(T)$ (as well as all the other components in the calculation). In practice, we will typically be more uncertain about the interpretation of our evidence than this procedure implies.

Going back to the example of equation 1, consider the possibility that we are uncertain about $\Pr(x = x_0|\neg T)$. Substantively, this corresponds to the idea that we aren't sure how often innocent people do not have alibis. In that case, we need to integrate over the distribution of this uncertainty, $g_x(q_x)$:

$$\Pr(T|x = x_0) = \int \frac{p_x \Pr(T)}{p_x \Pr(T) + q_x(1 - \Pr(T))} g_x(q_x) dq_x$$

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_x, the probability that guilty people do not have an alibi. In this case, we must determine:

$$\Pr(T|x = x_0) = \iint \left(\frac{p_x \Pr(T)}{p_x \Pr(T) + q_x(1 - \Pr(T))} f_x(p_x|q_x) \right) g_x(q_x) dp_x dq_x \tag{2}$$

This is our posterior belief about the probability that the suspect is guilty. However, it is important to point out that probability 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:

$$h(t|x = x_0) = \iint_{S(t)} f_x(p_x|q_x) g_x(q_x) dp_x dq_x$$
$$S(t) = \left\{ (p_x, q_x) \middle| \frac{p_x \Pr(T)}{p_x \Pr(T) + q_x(1 - \Pr(T))} = t \right\}$$

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 calcu-

lation to accomplish them all (Robert and Casella, 2004). In the example above, as long as the distributions $f_x(\bullet)$ and $g_x(\bullet)$ 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 2 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.

This procedure can be extended to uncertainty about other parameters as well, including our prior belief about the theory. Consider the previous example of forming a conclusion about the guilt of a suspect who has a motive and no alibi:

$$\Pr(T|x = x_0, y = y_1) = \frac{p_x p_y \Pr(T)}{p_x p_y \Pr(T) + q_x q_y (1 - \Pr(T))} \tag{3}$$

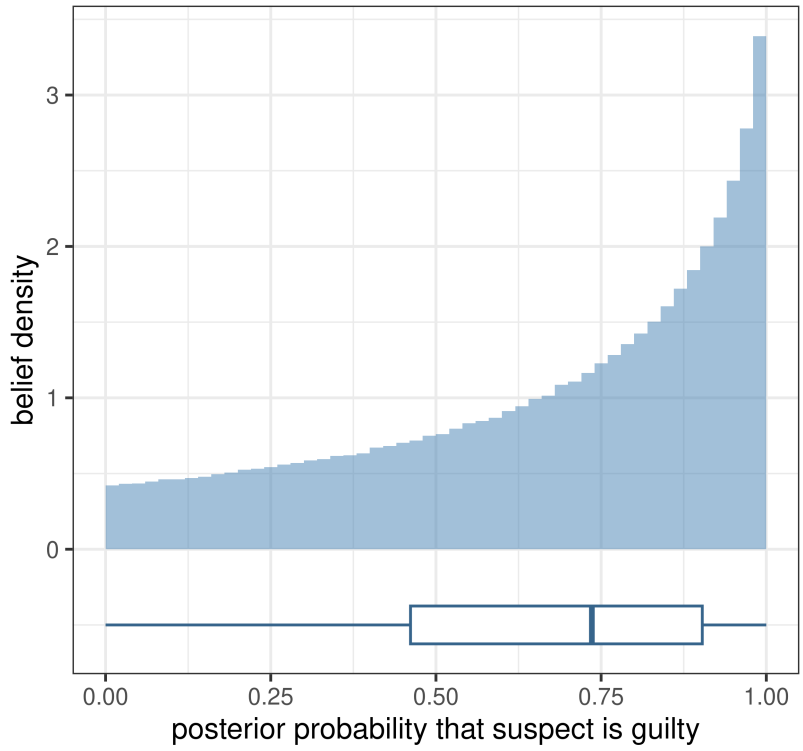
As laid out above, there are five pieces of information that must be supplied and about which we might be uncertain: p_x, q_x, p_y, q_y, and the prior $\Pr(T)$. As probabilities, these are bounded between 0 and 1 and any distributions must be similarly bounded. We therefore use truncated normal distributions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

$$p_x \sim \phi_T(0.95, 0.1) \qquad q_x \sim \phi_T(0.5, 0.2)$$
$$p_y \sim \phi_T(0.9, 0.1) \qquad q_y \sim \phi_T(0.7, 0.25)$$
$$\Pr(T) \sim U[0, 1]$$

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 3. We produced this belief distribution by drawing one million samples from

Figure 3: Mapping uncertain evidence into conclusions



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 (46 out of 100) 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.

Confidence and likelihood

The distinction between confidence and probability (or likelihood) raised by the previous example is worth lingering on for a moment. In some ways, the laws of probability obscure this distinction because in some cases it does not matter. Consider, for example, an outcome that has value v_a if event A occurs and value v_0 otherwise. Then the expected

outcome is:

$$\Pr(A) * v_a + (1 - \Pr(A))v_0 \tag{4}$$

In this case, it does not matter if A is actually a composite of multiple possibilities. For example, suppose that there are two ways that A might occur, one more likely than the other, that depends on the presence of an *a priori* unknowable stochastic background factor x. Then:

$$\begin{aligned}\Pr(x)\,[\Pr(A|x) * v_a + (1 - \Pr(A|x)) * v_0] + \\ (1 - \Pr(x))\,[\Pr(A|\neg x) * v_a + (1 - \Pr(A|\neg x)) * v_0]\end{aligned}$$

We can rearrange this as:

$$\begin{aligned}[\Pr(x)\Pr(A|x) + (1 - \Pr(x))\Pr(A|\neg x)]\,v_a + \\ [\Pr(x)(1 - \Pr(A|x)) + (1 - \Pr(x))(1 - \Pr(A|\neg x))]\,v_0\end{aligned} \tag{5}$$

Equations 4 and 5 are equivalent because of the principle of additivity:

$$\begin{aligned}\Pr(x) * \Pr(A|x) + (1 - \Pr(x)) * \Pr(A|\neg x) = \\ \Pr(A|\neg x) + \Pr(x)(\Pr(A|x) - \Pr(A|\neg x)) = \Pr(A)\end{aligned}$$

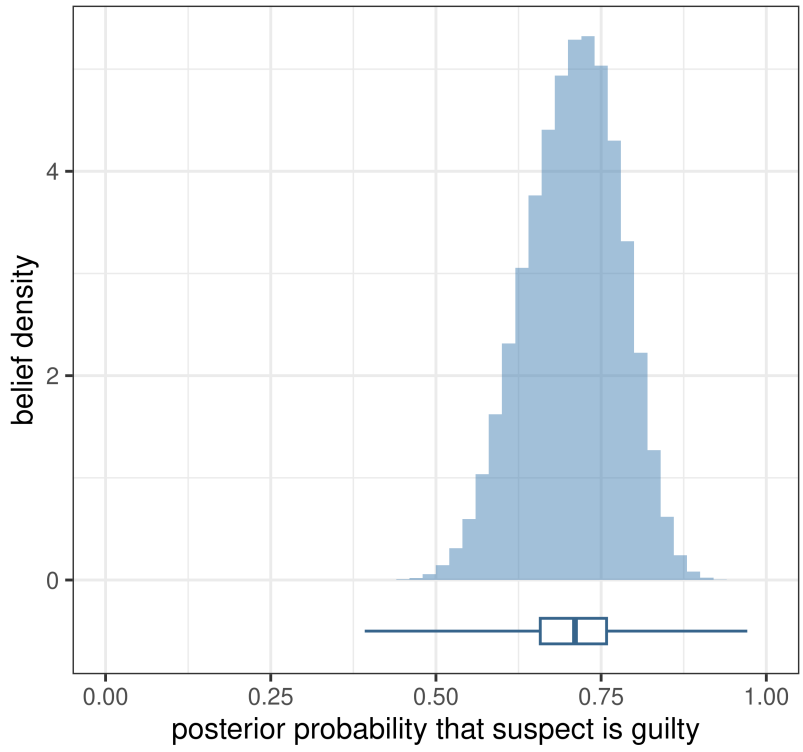
The equivalence of equations 4 and 5 is the core of the expected utility theorem describing rational choice under uncertainty (Mas-Colell, Whinston and Green, 1995, pp. 176-178). It is a property of von Neumann-Morganstern utility functions, which encapsulate patterns of preference over uncertain outcomes under a small set of assumptions designed to encapsulate rationality. The reducibility of compound lotteries is so foundational to rational choice theory that it will often be perceived as intuitive to scholars in that area; it is written indelibly on the heart of anyone trained in the rational choice tradition.

And yet, if our purpose is to come to accurate conclusions about the probability that a theory is true, compound lotteries are not reducible in this way. Consider the following revision of our continuing example, where we have revised our assumptions about the evidence of a suspect's guilt:

$$p_x \sim \phi_T(0.95, 0.05) \qquad q_x \sim \phi_T(0.5, 0.1)$$
$$p_y \sim \phi_T(0.9, 0.05) \qquad q_y \sim \phi_T(0.7, 0.1)$$
$$\Pr(T) \sim U[0.4, 0.6]$$

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.

Figure 4: A more confident belief in the same probability of guilt



These revised assumptions translate into the same expected belief as before: 71 out of every 100 suspects with a motive and no alibi will be guilty. But our certainty in this belief is much stronger, as shown by Figure 4. Now, we think there is only a very small chance (0.17%) that fewer than half of suspects are guilty under these conditions.

What is the practical difference in the beliefs represented by Figure 3 and Figure 4? Both say there is an $\approx 74\%$ chance that a theory is true—that we are "right," if our goal is to provide evidence for that theory—and thus an $\approx 26\%$ chance that we are wrong. But Figure 3 illustrates that there is a chance that we are *very* wrong in a way that isn't as possible for the beliefs in Figure 4.

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 3 or Figure 4, 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:

$$L(t, t^\star) = (t - t^\star)^2 \tag{6}$$

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 6), 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:

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

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 4 as being different than those of Figure 3, and to prefer the more confident beliefs from Figure 4.

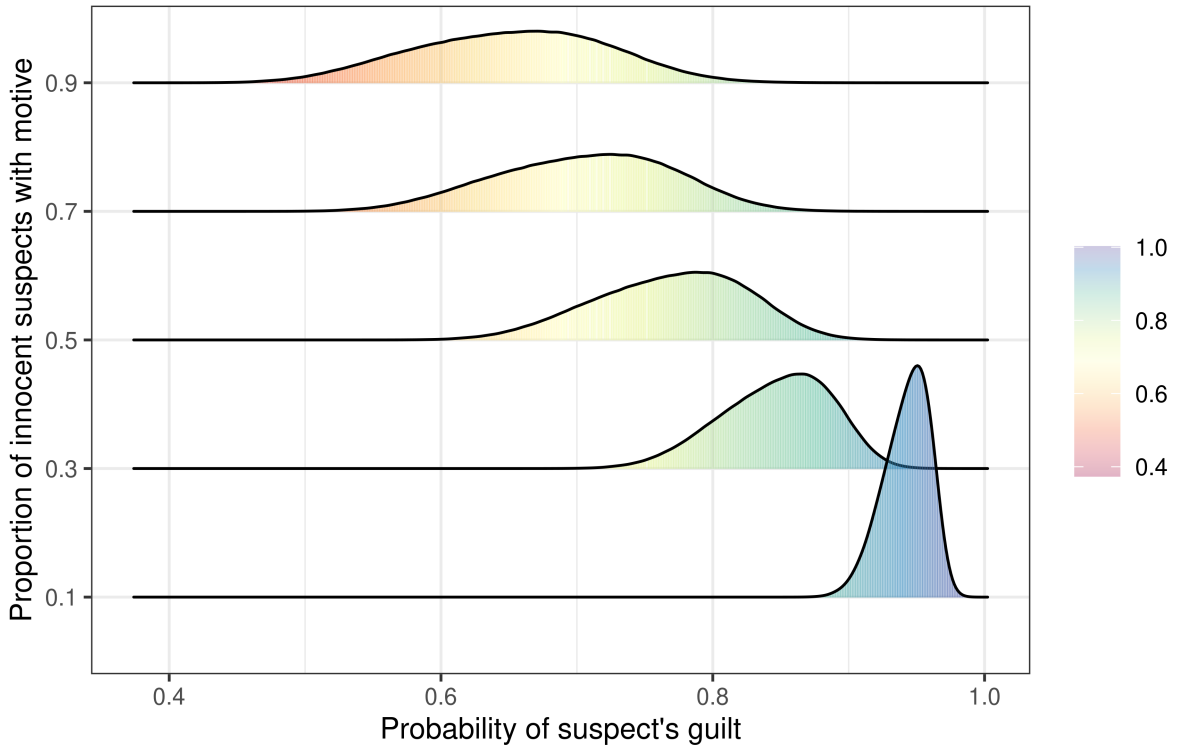
Robustness checking

The example above shows how more restrictive prior beliefs about a suspect's guilt translates into different (and more certain) posterior beliefs about that guilt given the same evidence. It also demonstrates the general principle of assessing the robustness of the conclusions of BPT to the assumptions that go into the process. These assumptions include prior beliefs, but also (and perhaps more importantly) include differences in interpretation of the evidence.

For example, one of the key pieces of evidence in the example is that the suspect had a motive to commit the crime. We assumed above that 70% of plausible but innocent suspects have a motive for committing the crime, with uncertainty about that interpretation. How sensitive is our conclusion to this interpretation of the evidence? To find out, we can systematically vary this assumption to see how it impacts our conclusions (holding all the other elements of the calculation constant).

Figure 5 demonstrates this idea. In this figure, we set all parameters in the same way

Figure 5: Assessing the robustness of conclusions to interpretation of evidence



we did for Figure 4 except for p_y. For this parameter, we set a fixed value of 0.1, 0.3, 0.5, 0.7, or 0.9 and calculate $\Pr(T|x = x_0, y = y_1)$ as before but separately for each of these values. We then plot those posterior belief distributions separately. As the Figure shows, our conclusion depends on how we think about the importance of a suspect having a motive. If we think that almost no innocent people have a motive ($p_y = 0.1$), we strongly believe in the suspect's guilt. But if we think even slightly more innocent people have motives, our posterior beliefs about this suspect's guilt become considerably less confident.

Testing partially rival and partially complementary hypotheses

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 hypothesis that one particular suspect is guilty is *partially rival* with the hypothesis that other suspects are guilty. Symmetrically, there are also *partially complementary hypotheses* where the truth of one will tend to be positively associated with the truth of the other.

As noted in the introduction, current Bayesian process tracing methods are unable to simultaneously test partially rival or partially complementary theories (Zaks, 2021, p. 67). If two theories are mutually exclusive—that is, if only one theory from the pair can be true—then Bayes' factors can be used to examine the degree to which evidence supports one theory over the other (Fairfield and Charman, 2017). A Bayes factor is the ratio of posterior probability densities for two theories, $\{T_1, T_2\}$. If there are two suspects and at most one can be guilty, then we should form beliefs about those two suspects' guilt given evidence x in the following way:

$$\Pr(T_1|x) = \frac{p_x \Pr(T_1)}{p_x \Pr(T_1) + q_x \Pr(T_2) + r_x \left(1 - \Pr(T_1) - \Pr(T_2)\right)}$$
$$\Pr(T_2|x) = \frac{q_x \Pr(T_2)}{p_x \Pr(T_1) + q_x \Pr(T_2) + r_x \left(1 - \Pr(T_1) - \Pr(T_2)\right)}$$

Where p_x is the likelihood of observing the evidence x if the first suspect is guilty, q_x is that likelihood if the second suspect is guilty, and r_x is that likelihood if neither is guilty. The Bayes factor comparing these two theories is:

$$BF = \frac{p_x \Pr(T_1)}{q_x \Pr(T_2)}$$

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, and is perhaps most useful when they are mutually exhaustive (that is, when one of the two theories *must* be true and the Bayes factor tells us which we should believe given the evidence).

Copula modeling of correlation among the truth of interdependent theories

But what about the situations shown in Figure 6? Here, we also have 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, although even when $|\rho| = 1$ the theories are not mutually exclusive unless $f(\tau_j|T_i)$ is a point mass at $\tau_j = 0$ for $i \neq j$. Thus, when we form a posterior belief about T_1, we must incorporate what we know about T_2 into this belief. Let τ_i equal the probability that T_i is true. We can write:

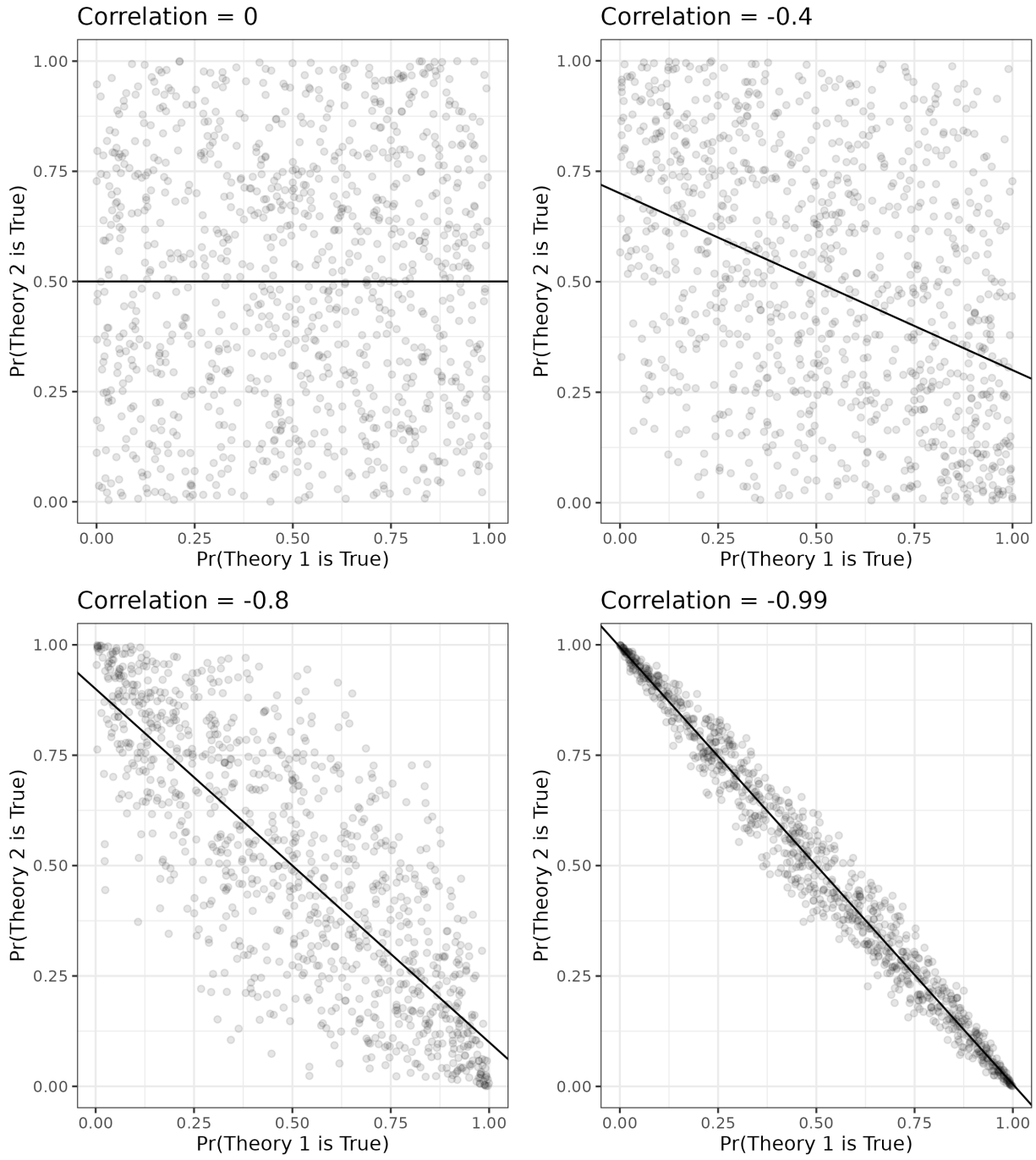
$$\begin{aligned}\Pr(T_1|x,\tau_2) &= \frac{\Pr(x,\tau_2|T_1)\Pr(T_1)}{\Pr(x,\tau_2|T_1)\Pr(T_1) + \Pr(x,\tau_2|\neg T_1)(1-\Pr(T_1))} \\ &= \frac{\Pr(x|\tau_2,T_1)f(\tau_2|T_1)\Pr(T_1)}{\Pr(x|\tau_2,T_1)f(\tau_2|T_1)\Pr(T_1) + \Pr(x|\tau_2,\neg T_1)f(\tau_2|\neg T_1)(1-\Pr(T_1))}\end{aligned}$$

where $f(\tau_2|T_1)$ is the probability distribution function for τ_2 when Theory 1 is true. Under most circumstances we can further simplify this expression to:

$$\Pr(T_1|x,\tau_2) = \frac{\Pr(x|T_1)f(\tau_2|T_1)\Pr(T_1)}{\Pr(x|T_1)f(\tau_2|T_1)\Pr(T_1) + \Pr(x|\neg T_1)f(\tau_2|\neg T_1)(1-\Pr(T_1))}$$

because the (possibly counterfactual) consistency of a piece of evidence under a particular theory would usually not be influenced by probability that a different theory is true.

Figure 6: Partially rival hypotheses



Conditional probability densities for interdependent theories

We presume that the researcher knows that $\Pr(T_1|x)$ and $\Pr(T_2|x)$ are correlated, but not the exact relationship between the two. We model this by assuming that they are distributed using a normal copula with the appropriate correlation. By approaching the problem this way, the marginal distributions of $\Pr(T_1|x)$ and $\Pr(T_2|x)$ are uniform; thus, the assumption of interdependence does not add information to our conclusions (unless we fix the value of one or the other). This distribution implies a probability density for $\Pr(T_i|x, T_j)$ and $\Pr(T_i|x, \neg T_j)$; for example, if T_j is true, then:

$$f\left(\tau_i|x, T_j\right) \propto \int_0^1 \left(f(\tau_i, \tau_j|x) \times \tau_j\right) d\tau_j \tag{7}$$

That is, we integrate out τ_j from the joint distribution at the target value of τ_i, weighted by the probability that T_j is true given the particular value of τ_j (which is simply τ_j). This function must be normalized by the total area of the function under τ_i to become a proper probability density (i.e., that sums to 1 under its admissible range). Equation 7 has a simple form: it is the least-squares regression line for the copula, the line shown in each of the examples of Figure 6, with slope equal to the correlation.

Drawing samples from interdependent posteriors

The conditional posterior belief densities implied by equation 7 require a more complex sampling strategy than our earlier examples. Drawing samples from the component densities to construct a sample, as we did in Figures 3 and 4, will no longer work because the likelihood of τ_1 depends on the value of τ_2 and vice versa.

Instead, we must employ Gibbs sampling (Robert and Casella, 2004, Chapter 9). The method is a familiar tool from Bayesian statistics. We construct 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:

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

That is, the value of τ_j upon which the distribution of τ_i depends are 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). The conditional distributions can be sampled from as before, by simply sampling from the component distributions.

An example inference

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:

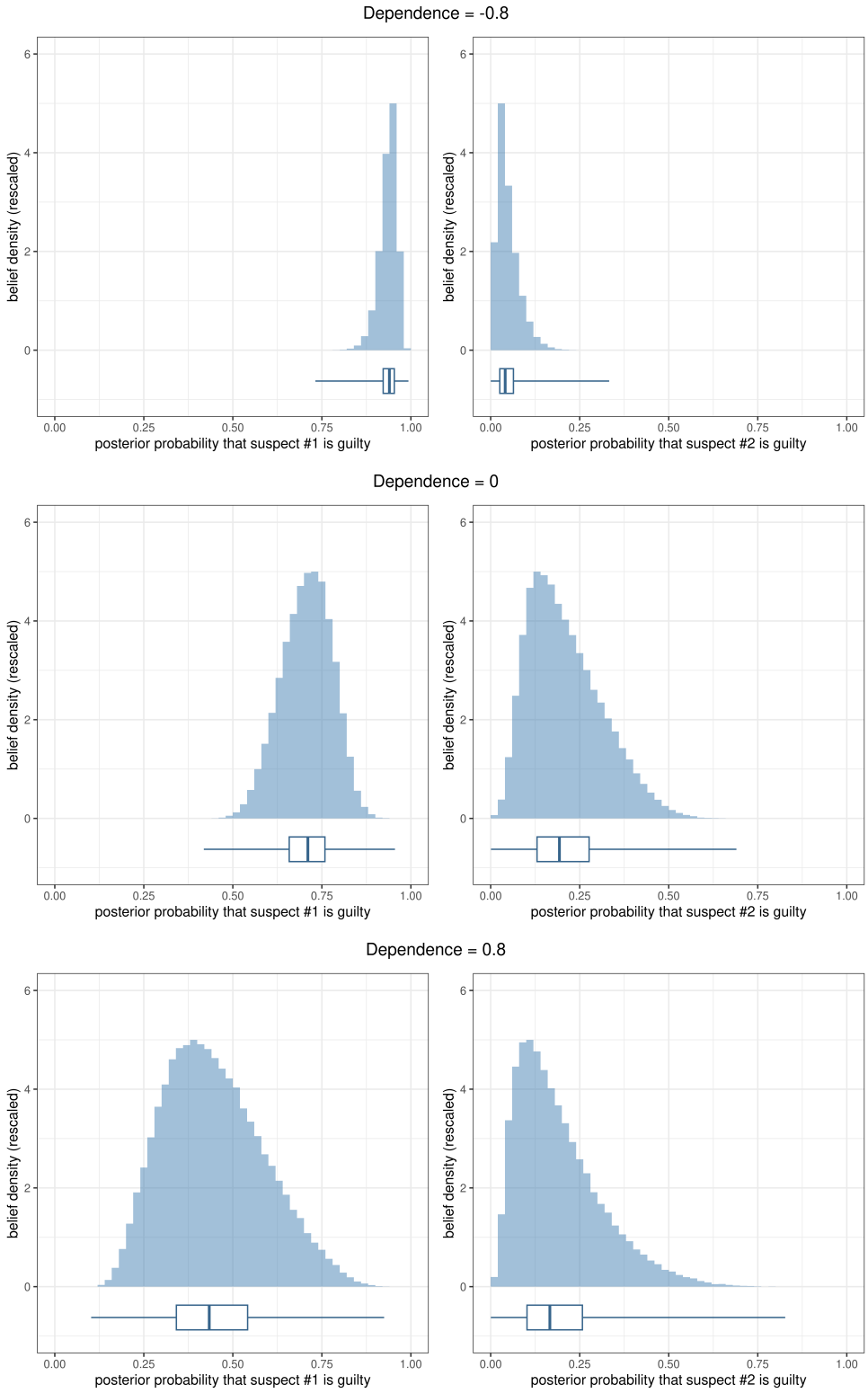
$$p_{1x} \sim \phi_T(0.95, 0.05) \qquad q_{1x} \sim \phi_T(0.5, 0.1)$$
$$p_{1y} \sim \phi_T(0.9, 0.05) \qquad q_{1y} \sim \phi_T(0.7, 0.1)$$
$$\Pr(T_1) \sim U[0.4, 0.6]$$

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

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.

Figure 7 shows the result of an analysis of this evidence presuming different values of correlation between the guilt of the two suspects. The middle panel, with dependence $\rho = 0$, is our posterior belief about suspect #1 (left) and suspect #2 (right) when the suspects' guilt is statistically independent; in this case, the evidence leads us to conclude that suspect #1 is likely but not indubitably guilty, whereas our beliefs about suspect #2 remain very uncertain. By contrast, if we know that the theories are partially rival (top panel, dependence $\rho = -0.8$), the evidence much more strongly incriminates suspect #1 and exonerates suspect #2. However, 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.

Figure 7: Jointly testing partially rival hypotheses



Software

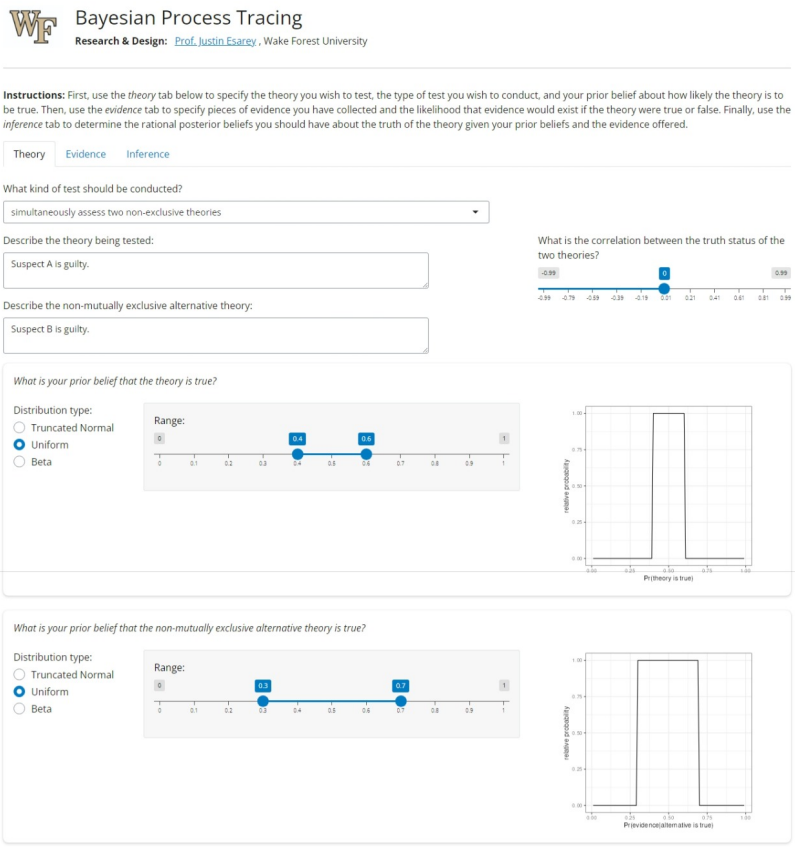
All of the analysis in this paper can be implemented in the R statistical software environment (R Core Team, 2024), and full replication code will be available upon publication of the paper. However, the paper also introduces a Bayesian Process Tracing application that also allows users to implement these methods without a need to code. That application is available at https://shiny.justinesarey.com/bpt-app. It uses the Shiny Server web hosting platform (Chang et al., 2024), which relies on R for the underlying computation and data visualization. The underlying for the Shiny application will also be available as part of the replication materials for this paper on publication.

A screenshot of the Shiny application is shown in Figure 8. The figure shows the initial screen for the application when the user loads it in a browser. The application allows for three different types of test to be conducted: comparing a theory to its logical complement (the theory being false, sometimes referred to as the null hypothesis); comparing a theory to a mutually exclusive and non-exhaustive rival hypothesis; and simultaneously assessing two non-exclusive theories with a possibly interdependent truth status. This screen 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).

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 tab. An arbitrary number of 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.

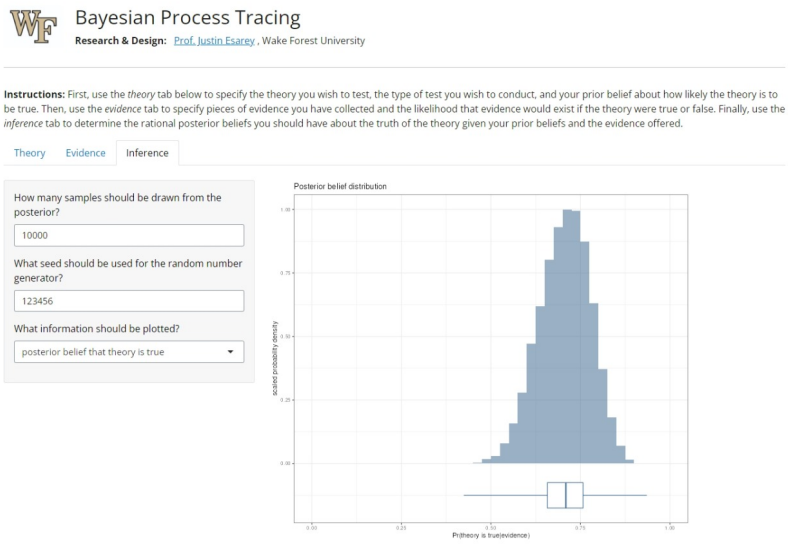
We inputted the evidence from the example in Figure 7 above, with suspect 1 having no alibi and a motive to commit a crime, and produced the posterior belief distribution

Figure 8: A screenshot of the initial screen of the Bayesian Process Tracing Shiny app.



shown in Figure 9 using the app. For this computation, we set $rho = 0$. Figure 9 shows the posterior belief density for the theory that Suspect 1 is guilty.

Figure 9: A screenshot of the inference screen of the Bayesian Process Tracing Shiny app for the crime problem shown in Figure 7, $\rho = 0$



This exercise using the web application makes it clear that inferences drawn using this procedure are highly transparent and easily replicable: the inferences from Figure 7 for $\rho = 0$ are substantively identical to those from Figure 9, even though Figure 7 was produced natively in R without use of the app and Figure **??** was produced much later. It also makes it clear how easy it is to check the robustness of conclusions drawn using this procedure: once we have successfully replicated the result in the app, we can adjust the parameters for the prior distributions and the likelihoods for pieces of evidence to determine how much the original conclusion changes as these assumptions change.

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 partially rival and/or 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

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.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges. 2024. *shiny: Web Application Framework for R*. R package version 1.8.1.1, https://github.com/rstudio/shiny.
URL: *https://shiny.posit.co/*

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

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.".

Fairfield, Tasha and Andrew Charman. 2019. "A dialogue with the data: The Bayesian foundations of iterative research in qualitative social science." *Perspectives on Politics* 17(1):154–167.

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

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

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

Mas-Colell, Andreu, Michael D. Whinston and Jerry R. Green. 1995. *Microeconomic theory*.

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

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

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.

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

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