A Appendix

This appendix presents a complete Bayesian process tracing analysis of the hypotheses and evidence laid out in the main article. I walk through my reasoning and uncertainty for each stage of the analysis, starting with hypothesis articulation and ending with inference.

Hypotheses & Evidence

Given two potential "causal factors" (i.e. murderers—Adam and Bertrand) Bennett, Fairfield, and Charman (2021), derive the following three hypotheses:

H_A: Adam committed the crime alone.

H_B: Bertrand committed the crime alone.

H_C: Bertrand lured the victim to the crime scene, Adam then committed the murder.

Drawing partly on Bennett, Fairfield, and Charman (2021), I examine three pieces of evidence.[1]

E_{car}: Eyewitness account of Adam's car near the scene of the crime on the night of the murder.

E_{rec}: Receipts placing Bertrand at a dim sum restaurant in the San Gabriel Valley at 7:45 PM on the night of the murder.

E_{vm}: A voicemail on the victim's phone from Bertrand requesting that the victim drop off documents near the crime scene by 7:30 PM on the night of the murder.

Assigning Priors & Likelihoods

The next step in the process is to assign the prior probabilities for each hypothesis as well as the likelihood—the probability of observing evidence E_i in world H_j. Following the authors' lead (and since I have no reason to believe otherwise) I assign equal priors to each of the three hypotheses. Thus, $P(H_A|I) = 0.33$, $P(H_B|I) = 0.33$, and $P(H_C|I) = 0.33$. The likelihoods for each piece of evidence I discuss in greater detail below. Where applicable, I follow the reasoning Bennett, Fairfield, and Charman lay out to inform the relative likelihoods.

E_{car}—First, I assign probabilities of finding evidence of Adam's car near the scene of the crime under each respective hypothesis.

[1] The authors only present the latter two pieces. I add the car because it is similar in type and I wanted to examine multiple rounds of the updating process.

$P(E_{\text{car}}|H_A) = 0.65$. Evidence of Adam's car near the scene of the crime would be relatively unsurprising in the world where Adam were the sole perpetrator. Depending on (1) how careless Adam is, (2) how common his car is, (3) how reliable the witness is, and (4) how frequently Adam is in that neighborhood, this evidence provides some, but not overwhelming support for the hypothesis.

$P(E_{\text{car}}|H_B) = 0.1$. As I mention in the main text, assigning a probability to this likelihood is considerably more difficult because the presence of Adam's car and Bertrand murdering someone are (in the world of H_B) independent events. As such, $P(E_{\text{car}}|H_B)$ reduces to $P(E_{\text{car}})$—or, more specifically, the frequency with which Adam drives in that area of Echo Park. An eyewitness account of Adam's car in Echo Park could be considerably higher if (1) Adam lives in or frequents Echo Park, and/or (2) Bertrand knows Adam lives there and intentionally choose the murder scene in an attempt to offset blame.

For the sake of ease, I assume Adam goes to Echo Park 1 in every 10 times he drives somewhere, which would place the probability of observing his car there on the night of the murder 0.1.

$P(E_{\text{car}}|H_C) = 0.65$. Given what we know, this evidence is equally likely under both H_A and H_C, and it does not help adjudicate among them, so I use the same value for both.

E_{rec}—The second piece of evidence is the receipt Bertrand's lawyer produced, placing him at a dim sum restaurant in the San Gabriel Valley at 7:45 on the night of the crime.

$P(E_{\text{rec}}|H_A) = 0.14$. Similarly to $P(E_{\text{car}}|H_B)$, Bertrand going to get dim sum and Adam committing a crime are independent events. Thus, the conditional probability reduces to the frequency with which Bertrand gets a hankering for steamed pork buns. Assuming the craving hits around once a week, the probability that Bertrand would happen to be out to dim sum on the night of the murder is 0.17.

$P(E_{\text{rec}}|H_B) = 0.005$. This quantity is also difficult to deal with. In BFC's letter, the authors maintain that credit card receipts placing the suspect out of town would be very unlikely if he committed the murder—and they go on to note that forgery of the receipts is too small of a probability to be worthy of consideration. However, if Bertrand is a savvy murderer, giving a friend, kid, or accomplice your credit card to create a credible alibi is quite reasonable.

While I stick with the authors' suggestion that this evidence is assessed as highly unlikely under H_B, this situation strikes me as an optimal one to suggest that researchers search for more evidence regarding Bertrand's habits, and the nature of the visit to the restaurant.

$P(E_{\text{rec}}|H_C) = 0.14$. The reasoning for this assessment is the same as that which guided $P(E_{\text{rec}}|H_A)$.

E_{vm}—Finally, the third piece of evidence for which we must assess likelihoods is finding a voicemail from Bertrand to the victim instructing them to drop of documents to a building near the scene of the crime by 7:30 PM.

$P(E_{\text{vm}}|H_A) = 0.05$. This probability is a quantity that depends on a variety of other factors, but for which I ultimately adopted the authors' logic for the sake of consistency. The probability of receiving a voicemail from Bertrand instructing the victim to drop of documents at the scene of the crime in the world where Adam was the sole murderer depends on (1) the victim's relationship with Bertrand, (2) the frequency with which the victim runs this (or this sort of) errand, (3) whether Adam knows of Bertrand (and the victim's relationship with him). Specifically, the voicemail and Adam's killing could be entirely independent. If Adam got into a fight with someone who happened to be in his way (i.e. the victim), then Bertrand's voicemail has nothing to do with his own guilt or Adam's, though it created a tragic circumstance for the victim. How do we assess the surprisingness of that situation? $P(E_{\text{vm}}|H_A)$ should just reduce to the frequency with which Bertrand asks the victim to run this errand.

Alternately, if the victim worked for Bertrand, consistently ran this errand, and Adam knew the victim and the victim's habits or schedule, then, once again, this evidence would be quite unsurprising, as it indicates that Adam is exploiting an errand to link Bertrand to the crime. Perhaps, then, the evidence and the hypothesis are not conditionally independent, but still unsurprising. Then the question becomes, are these different hypotheses?

$P(E_{\text{vm}}|H_B) = 0.1$. Following Bennett, Fairfield, and Charman's lead, I assign a fairly low probability to the likelihood that Bertrand both left a voicemail on the victim's phone asking them to go to the crime scene and also being the sole murderer. As they note, this decision is exceedingly bad for even the most dim-witted criminal.

Evaluating this probability, however, raises questions about how the sequence of evaluation (and confidence in interpretation) conditions our assignments of probability. Specifically, if we take the previous piece of evidence at face value—indicating that Bertrand was indeed out of town at the time of the murder—this piece of evidence seems more out of place in the world in which Bertrand were the sole murderer. If, however, we learned this piece of evidence before the lawyer produced receipts, I have a hard time believing researchers would immediately assign such a low probability to this likelihood—careless as Bertrand may be.

$P(E_{\text{vm}}|H_C) = 0.6$. Following the authors' lead, I assign a relatively high probability to this likelihood, as it would be unsurprising under the hypothesis specifying that Bertrand did the luring, while Adam did the murdering.

That said, I do not know why Bertrand leaving a voicemail on the victim's phone as part of a planned murder with an accomplice is any less careless than if Bertrand were the sole killer. To me, this piece of evidence suggests the need to find more evidence uncovering the relationship between Bertrand and the victim. Once again,

we find ourselves in a situation in which the answer is "it depends." If, again, Bertrand knows the victim and often sends them on errands, then perhaps this piece of evidence is more likely. If the document drop is out of the ordinary, then the voicemail is nearly as careless in this world as it is in the previous one.

Updating & Inference

Given the prior probabilities and likelihoods assigned above, I assume the next step is to conduct pairwise comparisons to assess the relative odds of the respective hypotheses using the following equation:

$$\frac{P(H_i|\mathbb{E}I)}{P(H_j|\mathbb{E}I)} = \frac{P(H_i|I)}{P(H_j|I)} \times \frac{P(E_1|H_iI)}{P(E_1|H_jI)} \times \ldots \times \frac{P(E_n|H_iI)}{P(E_n|H_jI)}. \tag{1}$$

The process at this point is not always clear as sometimes the previous literature jumps straight into assessing just the likelihood ratio for a single piece of evidence. Yet, the letter informs us that we can just use this equation to compute the odds in one step (which, to be sure, follows logically, as updating is built into the multiplication). As such, I first compute the relative odds of H_A (Adam committing the murder alone) and H_B (Bertrand committing the murder alone).

$$\begin{aligned}
\frac{P(H_A|\mathbb{E}I)}{P(H_B|\mathbb{E}I)} &= \frac{P(H_A|I)}{P(H_B|I)} \times \frac{P(E_{\text{car}}|H_AI)}{P(E_{\text{car}}|H_BI)} \times \frac{P(E_{\text{rec}}|H_AI)}{P(E_{\text{rec}}|H_BI)} \times \frac{P(E_{\text{vm}}|H_AI)}{P(E_{\text{vm}}|H_BI)} \\
\frac{P(H_A|\mathbb{E}I)}{P(H_B|\mathbb{E}I)} &= \frac{0.33}{0.33} \times \frac{0.65}{0.1} \times \frac{0.14}{0.005} \times \frac{0.05}{0.1} \\
\frac{P(H_A|\mathbb{E}I)}{P(H_B|\mathbb{E}I)} &= \frac{0.0015}{0.0000165} = 91
\end{aligned}$$

The result suggests that Adam committing the murder alone is 91 times more likely than Bertrand having committed the murder alone, given the evidence. Next, I compare the relative odds of H_A and H_C (Adam and Bertrand colluding).

$$\begin{aligned}
\frac{P(H_A|\mathbb{E}I)}{P(H_C|\mathbb{E}I)} &= \frac{P(H_A|I)}{P(H_C|I)} \times \frac{P(E_{\text{car}}|H_AI)}{P(E_{\text{car}}|H_CI)} \times \frac{P(E_{\text{rec}}|H_AI)}{P(E_{\text{rec}}|H_CI)} \times \frac{P(E_{\text{vm}}|H_AI)}{P(E_{\text{vm}}|H_CI)} \\
\frac{P(H_A|\mathbb{E}I)}{P(H_C|\mathbb{E}I)} &= \frac{0.33}{0.33} \times \frac{0.65}{0.65} \times \frac{0.14}{0.14} \times \frac{0.05}{0.6} \\
\frac{P(H_A|\mathbb{E}I)}{P(H_C|\mathbb{E}I)} &= \frac{0.0015}{0.018} = 0.083
\end{aligned}$$

This result suggests that the collusion is considerably more likely than Adam acting alone. By flipping the odds ratio (which I just find a more intuitive interpretation), H_C is 12

times more likely that H_A. Finally, I conduct the same analysis on H_C and H_B.

$$\frac{P(H_C|\mathbb{E}I)}{P(H_B|\mathbb{E}I)} = \frac{P(H_C|I)}{P(H_B|I)} \times \frac{P(E_{\text{car}}|H_CI)}{P(E_{\text{car}}|H_BI)} \times \frac{P(E_{\text{rec}}|H_CI)}{P(E_{\text{rec}}|H_BI)} \times \frac{P(E_{\text{vm}}|H_CI)}{P(E_{\text{vm}}|H_BI)}$$
$$\frac{P(H_C|\mathbb{E}I)}{P(H_B|\mathbb{E}I)} = \frac{0.33}{0.33} \times \frac{0.65}{0.1} \times \frac{0.14}{0.005} \times \frac{0.6}{0.1}$$
$$\frac{P(H_C|\mathbb{E}I)}{P(H_B|\mathbb{E}I)} = \frac{0.018}{0.0000165} = 1092$$

The final odds ratio suggests that H_C is 1092 times more likely than H_B, based on the evidence.

A point of unresolved disagreement is whether adding more hypotheses to the mix results in a combinatorics problem. If the goal is comparison across all hypotheses, then for n hypotheses, we will need to conduct $C(n, 2)$ computations. Yet, the authors of the letter maintain that "for n hypotheses there are only $(n-1)$ independent likelihood ratios" (7). They argue that as long as we have compared H_1 "to each of its $(n-1)$ rivals," we can use the following equation to "calculate likelihood ratios for all other pairs" (7):

$$\frac{P(E|H_jI)}{P(E|H_kI)} = \frac{\frac{P(E|H_jI)}{P(E|H_1I)}}{\frac{P(E|H_kI)}{P(E|H_1I)}}. \tag{2}$$

This instruction gives rise to both practical and inferential issues. First, if we have assigned likelihoods by "mentally inhabiting the world of each hypothesis," to assess the probability of finding the evidence, why would we want to do this step involving a reference hypotheses when we could just take the probability of observing E under H_j and pop it over $P(E|H_k)$ to compute the likelihood ratio? Moreover, using this equation is still going to require that we do more than $(n-1)$ computations—it might just be that the computations look different. Finally, this equation only gives the likelihood ratio of H_j to H_k, we can *only* use this equation on its own to assess H_j versus H_k if the researcher has uninformative (i.e. equal) priors, which again, the literature does not make clear.

A.1 The Narrative Approach

Throughout the BPT literature, its proponents often note that we need not always use the mathematical (filling in the numbers) approach. Instead, they argue, we should just take cues from Bayesian mathematics to reason through our evidence with a narrative comparison of likelihoods. Here, I discuss the process and subsequent questions that arise from using the narrative approach. While the original plan was to conduct a narrative analysis of the same evidence, the process ultimately raised so many questions and concerns that I opt instead to draw on the example to illustrate these problems, but I did not feel comfortable implementing an analysis and calling it Bayesian inference.

A.1.1 When should we use it?

Bennett argues that we need not use the mathematical approach for every piece of evidence, and that sometimes it makes sense to use the narrative approach in conjunction with computations, and to reserve the explicit mathematical tack "for the few pieces of evidence that a researcher considers the most probative" (Bennett 2014, 51). Similarly, Fairfield and Charman will note explicitly that they choose one over the other in different applications. I remain unclear, however, on how practioners should know how to adjudicate. How can we assess the "probative value" of evidence a priori? Once we do, what is the cutoff at which something is deemed sufficiently probative as to warrant the mathematical approach? Is there ever a clear benefit (or drawback?) to using one over the other? If so, how can we identify whether we're at that line?

A.1.2 Assigning Priors and Likelihoods

Assuming we do have reason to choose the narrative approach, the next step is to assign priors and likelihoods. Dealing with priors is easy enough when we have no reason to believe one hypothesis is any more likely that the others. However, I am left wondering how to deal with priors in the narrative BPT approach if we have reason to believe (or doubt) H_A over H_B. At the outset, it is easy enough to say "H_A seems considerably more plausible than H_B," but complications arise at the inferential stage.

Assigning likelihoods in the narrative approach is a more elusive process in which a number of problems arise. Again, the goal of the likelihood function is to "mentally inhabit the world of each hypothesis" to assess the probability of finding a given piece of evidence in that world (Fairfield & Charman 2019, 159). In short, we must ask *"if H_A is the true world, how surprised am I to find E_{car}?"* and so on for each piece of evidence and each hypothesis. The first issue, as I mention in the main text, is that researchers will have to decide on a ranking of language to indicate the likelihood of finding a given piece of evidence in a given world. In a project with more than three pieces of evidence (let alone more than three hypotheses), likelihood functions and subsequent updating could easily become confusing and intractable.

The second issue with likelihoods in the narrative approach is that the BPT literature exhibits a disconnect between what the likelihood function is supposed to be versus how it is assessed in examples. Specifically, the step of "mentally inhabiting" the world of a hypothesis to assess the likelihood of finding evidence there should be analytically prior to asking "is E_i more expected under H_j or H_k?" Jumping straight to the latter question can blind researchers to the host of questions that may arise when considering a hypothesis on its own.

For example, evidence of Adam's car near the scene of the crime may seem much more likely under H_A than H_B when thinking about them together. Yet, by asking "Should I expect to see Adam's car near the scene of the crime of Bertrand is the sole murderer?"—and truly inhabiting that world—we might then ask a host of follow-up questions that should guide further research. Does Adam live nearby? Is there a chance Bertrand knows this information and chose the murder scene intentionally to pin it on

Adam? Is Adam's car an especially popular make and model? Or, is it so rare that it would be foolish of Adam to be driving it with the intention to commit a crime? Deeply inhabiting the world of each hypothesis—rather than jumping straight to comparison—raises numerous questions revealing our uncertainty, which in turn should guide the search for more evidence. If ever there were a benefit to Bayesian process tracing, this would be it: a methodical process that guides us in the specific type of evidence we should collect.

By jumping straight into comparison, the narrative approach to likelihood assessment seems to ask researchers to condition probabilities on the rest of the hypothesis set, because we are essentially asking "how likely am I to see Adam's car if Bertrand is the sole murderer knowing that one of the other options is that Adam is the sole murderer?" This approach will likely lead researchers to systematically underestimate how "expected" evidence is under different hypotheses by blinding them to the various contingencies. It is only at the point of inference where we compare the likelihood of one hypothesis to the next.

Furthermore, by asking "whether and to what degree does the evidence fit *better* with that hypothesis as compared to rivals" (Bennett, Fairfield & Charman 2021, 5), this approach could inadvertently lower the standards for the quality of evidence needed to draw inferences. If all that matters is comparison, rather than whether and to what extent evidence supports a specific hypothesis, evidence that is weakly expected under H_A but completely unrelated to H_B is going to nonetheless bolster our confidence in the former. Many pieces of comparable evidence might lead researchers to strongly infer in H_A's superiority over H_B, despite having no evidence that H_B is incorrect, and very weak evidence of H_A's role.

Once again, the central problem is not that Bayesian process tracing does not work in theory, it is that aspiring Bayesians need to be equipped with a better sense of where they could falter in practice.

A.1.3 Updating & Inference

The process of Bayesian inference entails weighting likelihoods by our priors and then iteratively updating as we move through the evidence. The mathematical approach has the convenience of multiplication, which, in essence, does the iterative updating (i.e. weighting subsequent likelihoods by posterior updates) for us. In the narrative approach, however, we cannot use a one-shot equation for each pair of hypotheses to analyze all the evidence in a single computation. This added complication raises a number of procedural questions.

1. How should researchers evaluate likelihood ratios if we do not have equal priors on the respective hypotheses? Is this an indication that we *should* use the mathematical approach?

2. If researchers do go in with equal priors, how do we then incorporate our updated posterior estimates into our evaluation of the subsequent likelihood ratio? What is the process of "weighting" hypotheses verbally?

3. Assuming there is some way to adequately update, how can researchers keep track of the ranking of our respective hypotheses at a given time? How can we translate that ranking into a cohesive narrative on the page?

I am also left wondering how reviewers should evaluate the inferences and likelihoods on which they are based.

A.2 Closing Notes

An unexpected result of applying BPT in practice is that I came away more skeptical of the narrative approach than I previously had. In my original piece, I expressed considerably more skepticism for the mathematical approach and maintained that the narrative approach provided an intuitive framework for conducting an iterative evaluation of evidence. Yet, after attempting to use both, I the practical difficulties outweigh the intuitive appeal. Most inimically, the narrative approach seems, at second glance, to compromise the analytic transparency BPT is supposed to contribute. At least with the mathematical approach, researchers will be forced to more explicitly justify their selection of probabilities and their approach to satisfying the critical assumptions on which the method relies. Though I am still skeptical that reviewers will have the tools or experience needed to critically evaluate these probabilities in the peer-review prorcess.

More broadly, the problems I address here arose in the context of the simplest example: solving a murder with a closed list of suspects. I have in the past decried the use of these examples due to equifinality and the added layers of complexity in social science theories (including causal factors operating at different levels of analysis) (Zaks 2017). We use them to center the method itself—in the same way that we teach statistics with much better data than we will often collect in practice. The question I am left with is, how do we move to more complex data in a more complex world when the simplest example raised unanswered questions at each stage?

References

Bennett, Andrew. 2014. "Process Tracing with Bayes: Moving Beyond the Criteria of Necessity and Sufficiency." *Qualitative and Multimethod Research* 12(1):46–51.

Bennett, Andrew, Tasha Fairfield & Andrew Charman. 2021. "Understanding Bayesianism." *Political Analysis* XX(XX):1–12.

Fairfield, Tasha & 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.

Zaks, Sherry. 2017. "Relationships Among Rivals (RAR): A Framework for Analyzing Contending Hypotheses in Process-Tracing." *Political Analysis* 25(3):344–362.