

Bayesian Reasoning in the Social Sciences

Transparency and authorship note

Rod Malpert | Revised October 19, 2026

This is an AI-assisted Centaur Research work. Rod Malpert originated and directed the project, selected its questions, examples and substantive judgments, evaluated and revised the text, and retains responsibility for the final version. ChatGPT (OpenAI) supplied substantial research, synthesis, calculation and drafting assistance.

AI systems are acknowledged tools, not accountable authors. This disclosure accompanies the FAQ when circulated. Cite as: Malpert, Rodney. 2026. Bayesian Reasoning in the Social Sciences. AI-assisted Centaur Research FAQ, revised October 19, 2026. RodMalpert.com.

Bayesian Reasoning in the Social Sciences

A friendly FAQ on priors, evidence, AI assistance, and why vague words cannot hide our assumptions

Rod Malpert | Revised October 19, 2026

Bayesian reasoning makes starting assumptions, evidence judgments and their consequences explicit. Its strength is inspectability and consistency, not an assurance that numerical judgments are accurate. This FAQ explains the arithmetic, shows how to test assumptions, and gives equal attention to uncertain likelihoods, selection, dependence and omitted alternatives.

AI can reduce the work of calculation and evidence bookkeeping. It can also produce persuasive but poorly founded models, invent sources, and anchor a researcher on its first suggested values. Expertise remains necessary to assess the evidence and assumptions; human control and agreement between AI systems are not independent validation.

Three ideas to take away

1. Make background assumptions explicit. A verbal judgment can contain unresolved expectations without encoding one unique prior. Neither an adjective nor an arbitrary percentage becomes reliable merely by being stated.
2. The arithmetic is accessible; the evidence model is difficult. AI can calculate an update, but the hardest historical judgments often concern how probable the evidence would be under competing explanations.
3. Keep responsibility and uncertainty visible. Record measured quantities, elicited judgments, AI suggestions and unresolved assumptions separately. When no defensible numerical range is available, suspend a numerical verdict.

Part 1. The basics in ten minutes

1. What is Bayesian reasoning?

It is a method for updating what you believe when new evidence arrives. It has three ingredients:

- The prior: how plausible a claim is before the new evidence, given what you already know.
- The likelihood ratio: how much more (or less) expected the evidence is if the claim is true than if it is false. Think of it as the strength and direction of the push the evidence gives.
- The posterior: your updated assessment after taking the evidence into account.

Today's posterior becomes tomorrow's prior when the next piece of evidence arrives. That cycle is what learning from evidence is. In social science the claim might be that a policy improves outcomes, that an allegation is accurate, or that one explanation best accounts for a historical record.

2. Why do Bayesians talk about odds instead of percentages?

Because odds turn updating into a single multiplication. Odds compare the chances for a claim with the chances against it. A 20% probability means 20 chances for and 80 against, which simplifies to 1 to 4. A 50% probability is 50 to 50, or 1 to 1.

Two conversions are all you need. To turn a probability into odds, compare it with what is left over: 20% becomes 20 to 80, or 1 to 4. To turn odds of a to b back into a probability, divide a by the total, $a + b$: odds of 3 to 4 become $3 \div 7$, about 43%.

Probability	Odds (for to against)	Probability	Odds (for to against)
10%	1 to 9	60%	3 to 2
20%	1 to 4	67%	2 to 1
25%	1 to 3	75%	3 to 1
33%	1 to 2	80%	4 to 1
50%	1 to 1	90%	9 to 1

3. Show me the key calculation, one step at a time.

Worked example: a 20% prior meets evidence with a likelihood ratio of 3

Step 1. Convert the prior to odds. 20% means 20 chances for and 80 against: odds of 1 to 4.

Step 2. Multiply the "for" side by the likelihood ratio. The evidence is three times as expected if the claim is true, so $1 \times 3 = 3$. The odds become 3 to 4.

Step 3. Convert back to a probability. 3 to 4 means 3 chances for out of $3 + 4 = 7$ in total. $3 \div 7 =$ about 43%.

The whole rule in one line: posterior odds = prior odds $\times$ likelihood ratio.

Why not simply multiply 20% by 3 and get 60%? Because probabilities cannot exceed 100%. Apply the same shortcut to a 50% prior and you would get an impossible 150%. Odds have no ceiling, so the multiplication always works, and the conversion back to a percentage automatically keeps the answer between 0% and 100%.

The same calculation with people instead of formulas. Imagine 100 cases like the one in front of you. With a 20% prior, the claim is true in 20 of them and false in 80. Suppose the evidence shows up in 60% of the true cases (12 cases) and in 20% of the false cases (16 cases). Its likelihood ratio is $60\% \div 20\% = 3$. Now look only at the 28 cases where the evidence appeared: the claim is true in 12 of them. $12 \div 28 =$ about 43%, exactly the answer the odds gave. Psychologists have found that

counting cases this way makes Bayesian reasoning far easier for most people to follow [6].

4. What does “three to one” evidence actually mean?

A likelihood ratio of 3 means the evidence is three times as probable if the claim is true as it is if the claim is false. It is the size of the push, not the destination. It does not mean the claim is 75% likely, that a witness is three times as reliable, or that the evidence is “three times as impressive.” Where you land depends on where you started: the same factor of 3 takes a 20% prior to 43% but takes a 50% prior to 75%.

5. What happens when a second piece of evidence points the other way?

You multiply again. Suppose a second observation is only half as expected if the claim is true (likelihood ratio 0.5), judged after the first observation is already known. Starting from odds of 3 to 4: $3 \times 0.5 = 1.5$, so the odds become 1.5 to 4. Convert back: $1.5 \div 5.5 =$ about 27%.

Both observations count, and neither erases the other. Their combined effect is $3 \times 0.5 = 1.5$, so overall the claim moved from 20% to 27%. This shows an important distinction: evidence can raise confidence in a claim without making it more likely than not. “The evidence supports H” is incomplete until we say both which way the evidence pushed and where the claim ended up.

6. What makes evidence strong?

Evidence is strong when it is much more expected under one explanation than under the alternatives. “Consistent with” is not the same as “discriminates in favor of.” A politician praising a popular policy is consistent with sincere belief, but it is also exactly what a politician would do while pandering. Because both explanations predict it, its likelihood ratio is close to 1 and it tells us little.

Conventional verbal scales do exist: Kass and Raftery give interpretive categories for Bayes factors [5]. These describe relative evidential support, not posterior probability or a universal dictionary for historical adjectives. The posterior effect still depends on starting odds. For example, begin at 20%, or odds of 1 to 4. Multiplying those odds gives:

Evidence factor	Updated odds	Updated probability
1	1 to 4	20%
1.1	1.1 to 4	22%
1.5	1.5 to 4	27%
2	2 to 4	33%
3	3 to 4	43%
10	10 to 4	71%

Part 2. Nobody opts out of priors

7. Can I escape a prior by refusing to put a number on it?

You can refuse to write down a number. You cannot, by refusing, remove the starting assessment from your argument. The moment you call an explanation unlikely, ordinary, extraordinary, or contrary to consensus before looking at the new evidence, you have given it a starting position, and that position determines how far the evidence must move you.

The verbal assessment need not hide one precise number; its boundaries may simply be unresolved. The point is about function: background plausibility does exactly the job a prior does. Leaving it unstated does not make it less influential. It only makes it harder to inspect.

A scholar may responsibly suspend judgment. But a scholar who announces that a claim is probably false while declining to say how plausible it was to begin with has not escaped the problem of priors. Part of the reasoning has simply gone undisclosed.

8. What is wrong with words like “unlikely” or “unclear”?

Nothing, if they communicate the uncertainty well enough for the purpose. Trouble begins when the conclusion depends on differences the words conceal. “Unlikely” might mean 1%, 10%, or 30%. Watch what the same evidence, with a likelihood ratio of 5, does to each reading:

“Unlikely” read as	Prior odds	× 5 gives	Posterior
1%	1 to 99	5 to 99	about 5%
10%	1 to 9	5 to 9	about 36%
30%	3 to 7	15 to 7	about 68%

One word, one piece of evidence, and the conclusion ranges from “still doubtful” to “more likely than not.” A reader of the verbal version cannot tell which argument is being made.

“Unclear” is even less specific. It might mean balanced evidence, scarce information, an ill-defined hypothesis, or disagreement among experts. Those are different problems, and none of them should be silently translated into 50%.

Words mean different numbers to different people

This is not a hypothetical worry. In a famous 1964 essay, Sherman Kent of the CIA described discovering that colleagues who had all approved the phrase “serious possibility” in an intelligence estimate read it as anything from roughly 20% to 80% [2]. Later research surveyed by Mosteller and Youtz found the same pattern for everyday words such as “likely,” “probable,” and “possible”: people agree on the rough order of the words but vary widely in the numbers they attach to them [3].

Fields that must communicate uncertainty to the public have responded by pinning words to numbers. The Intergovernmental Panel on Climate Change, for example, defines “likely”

as 66 to 100%, “very likely” as 90 to 100%, and “unlikely” as 0 to 33% [4]. Social scientists can do the same, and Bayesian reasoning gives them a natural way to do it.

9. The same is true of weighing evidence, isn’t it?

Yes. When “compelling” or “suggestive” means that an observation favors one explanation over another, it asserts a comparison of how well each explanation predicts that observation, which is exactly what a likelihood ratio states. These words sometimes describe something else, such as overall confidence or source quality, so say which meaning is intended. An explicit comparison can be tested and revised; an adjective cannot.

10. Is 50% the right prior when I know nothing?

Not automatically. Two named possibilities do not become equally likely merely because there are two of them. “The policy works” and “the policy does not work” can each hide many effect sizes and mechanisms. A 50% prior needs a justification just as a 20% prior does. Sometimes the responsible choice is a broad range, or a decision not to estimate until the question is better defined.

Part 3. Building a prior, and disagreeing about one

11. Where can a reasonable prior come from?

- Comparable cases can provide base rates: how often have similar policies, claims, or sources turned out to be right?
- Earlier studies can inform a new analysis.
- Knowledge of mechanisms can rule out implausible effect sizes.
- Structured expert elicitation can reveal what specialists expect and why.
- Hierarchical models can borrow strength from related populations while allowing real differences.
- Weakly informative priors can exclude absurd extremes without dictating the answer.

None of these guarantees a correct prior. All of them make the prior explainable, which is the point. A useful habit is to ask what observations your prior and model predict together, and whether those predictions look sensible before the new evidence is used.

12. How should a comparison group (reference class) be chosen?

Choose comparable cases by similarities that matter to the question, and where possible fix the inclusion rules before seeing the result. A group that is easy to count is not automatically the right group. Selection bias, uncertain classifications, small samples, and differences between the group and the case at hand can all weaken the comparison.

Ask why membership in the chosen group should predict the disputed property of this case. Try alternative groups and allow uncertainty in their rates. If the conclusion swings sharply when the group changes, the disagreement is about a substantive assumption, not

about arithmetic, and that is worth knowing.

13. Does the prior have to be accurate to the percentage point?

No. A prior can be useful without settling whether the best value is 19%, 20%, or 21%. Use a range that reflects your real uncertainty, then ask whether the conclusion survives across that range. If every prior from 10% to 30% leaves a claim below 40% after the evidence, arguing over 19% versus 20% accomplishes little. If the plausible range puts the answer on both sides of 50%, the honest conclusion is that the result depends on that assumption, and now everyone can see exactly why.

A defensible range can reveal sensitivity, but stating a range does not establish its calibration. A careful qualitative assessment remains legitimate where the evidence cannot support useful numerical bounds.

14. Can reasonable people start with different priors?

Yes. Different justified priors can lead to the same practical conclusion or expose unresolved disagreement. If the conclusion changes across reasonable starting assessments, report that dependence. It may suggest useful research, but does not mean we know exactly which new observation would resolve the dispute.

15. Isn't putting a number on a judgment false precision?

Only when the presentation claims more certainty than the inputs support. "About 20%, with a defensible range of 10 to 30%" does not claim to know the exact percentage. An unexplained "20.137%" would. A number states an assumption more precisely; it does not claim the assumption is precisely known. Round the public conclusion, keep the full calculation available, and label a sensitivity range as a sensitivity range rather than a statistical interval.

Part 4. The same evidence in words and in numbers

The next puzzle illustrates updating, not the superiority of numerical conclusions. It combines an empirically motivated marginal comparison with two stipulated teaching ratios. That mixture must not lend the stipulated values borrowed credibility. Every factor would require population-specific, conditional justification in a real investigation.

The puzzle

The question is whether Pat is the man. The starting probability is one in three. The evidence is red-green vision deficiency, an education major, and a preference for women's volleyball. The ratios below are teaching assumptions; no claim is made that they are validated for this particular population or jointly conditional on these clues.

The worked comparison

Step	Verbal account	Odds and probability
Start	One of three is a man.	1 to 2; 33%
Red-green vision	Strongly favors man.	$\times 17 = 17$ to 2; 89%
Education major	Points toward woman.	$\times 0.5 = 8.5$ to 2; 81%
Women's volleyball	Points toward woman.	$\times 0.4 = 3.4$ to 2; 63%
Conclusion	Probably the man.	About 63%, given the inputs.

The factor of 17 is motivated by a comparison of inherited red-green vision-deficiency rates reported for people with Northern European ancestry [7]. That marginal comparison does not automatically supply the right conditional likelihood ratio for this selected group. The example does not establish a diagnosis or a real-world identification probability.

The education and sport ratios, 0.5 and 0.4, are entirely stipulated teaching values. They are not measured frequencies. All three ratios must be treated as conditional on earlier information; gender, education and sports preferences need not be independent. The illustration shows arithmetic under assumed inputs, not evidential calibration.

The calculation: $1 \text{ to } 2 \times 17 \times 0.5 \times 0.4 = 3.4 \text{ to } 2$, or about 63%. Each later clue is assessed given the earlier clues; simply multiplying marginal rates can double-count related evidence. The answer is conditional on the puzzle and stated weights.

16. What happens if we change the prior?

Keep the three factors fixed. Their combined effect is $17 \times 0.5 \times 0.4 = 3.4$. Different starting points produce different answers:

Prior	Prior odds	After $\times 3.4$	Posterior
20%	1 to 4	3.4 to 4	46%
1/3	1 to 2	3.4 to 2	63%
50%	1 to 1	3.4 to 1	77%

The numbers expose the starting assumption and let readers substitute one of their own.

17. What if the sport clue has a different weight?

After the color-vision and major clues, the odds are 8.5 to 2. Vary only the sport factor:

Sport factor	New odds	Posterior
0.25	2.125 to 2	52%
0.4	3.4 to 2	63%
0.6	5.1 to 2	72%

The exact answer depends on how informative that preference really is. A reader can replace 0.4 with a defensible estimate and see the effect immediately. The example does not require agreement about stereotypes.

18. What does the comparison establish?

It shows what the stated assumptions imply. The verbal verdict "probably the man" agrees with the numerical 63%, so the example does not demonstrate greater accuracy than careful qualitative reasoning. Numbers add a reproducible update and explicit sensitivity; neither validates the inputs.

Part 5. The AI opportunity: expert judgment, machine bookkeeping

19. Why is now the moment for Bayesian social science?

AI reduces the labor of converting odds, organizing evidence and calculating sensitivity cases. That makes explicit models accessible to more historians and social scientists. It does not resolve the substantive difficulty of specifying what the evidence predicts under competing explanations. The opportunity is easier construction and scrutiny, accompanied by a risk that weak inputs acquire a misleading numerical authority.

AI can free time for substantive review, but can also obscure it behind polished output. Expertise does not become more effective automatically. Ask the researcher to record an initial assessment before seeing AI-proposed numbers where feasible, require source checks, and examine alternatives that challenge the preferred result.

20. Does the scholar have to hand over control of the judgments?

No. There are several comfortable ways to work, and the researcher chooses:

Mode	Who sets priors and evidence weights	What the AI does
Expert in full control	The researcher, entirely from professional knowledge	Organizes the model, calculates, flags technical problems
Mode	Who sets priors and evidence weights	What the AI does
AI proposes, expert decides	The researcher, after reviewing suggested values	Offers candidate priors and weights with reasons; the expert accepts, adjusts, or rejects
Working dialogue	Built together through critical discussion	Argues alternatives, plays devil's advocate, tests the expert's reasoning

A useful standing instruction

“Keep my stated inputs fixed. Flag possible problems, explain why they matter, and show any proposed alternatives separately. Do not silently change my assumptions.”

The resulting record should distinguish human decisions, AI suggestions, and calculations. The scholar keeps responsibility for the final claims.

21. What can AI contribute beyond doing the math?

Quite a lot. An assistant can help build an evidence ledger, trace each observation back to its sources, notice when two arguments rest on the same premise, propose which probabilities should be conditional on which, separate competing explanations from stages of a single explanation, and test alternative structures. This makes explicit analysis available to domain experts who have not mastered every formal technique, provided the assistant explains its work clearly enough for the expert to review it.

22. How does it help catch double counting?

Double counting is one of the most common and most consequential errors in any argument, verbal or numerical. The test is simple to state: does this item add new information once the earlier items are known? Different labels or different authors do not make evidence independent. Two witnesses may repeat one rumor; a review article may summarize the very studies already counted; a prior may already include the finding later presented as new support.

Worked example: the echo effect

A report has a likelihood ratio of 3. A newspaper then reprints it without adding any new information.

Counted correctly: the reprint adds a factor of 1 (no new information). Starting from a 20% prior: odds 1 to 4, $\times 3 = 3$ to 4, which is about 43%.

Counted twice: 1 to 4 , $\times 3 \times 3 = 9$ to 4 , which is $9 \div 13$, about 69%.

One repeated source turned a doubtful claim into an apparently probable one. A verbal argument makes the identical error whenever it treats repeated coverage as mounting confirmation. The difference is that in an explicit ledger the error can be found.

Partial dependence is harder. It does not follow that the second item adds nothing, or that an arbitrary percentage discount fixes the problem. AI can help estimate conditional ratios, group overlapping evidence, and show how much the conclusion depends on the dependence assumption. The write-up should say which approach was used and what remains uncertain.

23. What mistakes involving branches can it help reveal?

Possibilities can be mutually exclusive, overlapping, or sequential, and each structure calls for a different calculation. Three quick illustrations:

- Overlapping possibilities. Mechanism A has a 40% chance, B has 30%, and both together 10%. The chance of A or B is $40\% + 30\% - 10\% = 60\%$, not 70%.

- Exclusive possibilities. Suppose A has probability 25% and B has probability 35%, and the two mechanisms are disjoint. If an outcome has conditional probabilities of 60% under A and 20% under B, their contributions sum to $0.25 \times 0.60 + 0.35 \times 0.20 = 22\%$. Other branches must be included if they can also produce the outcome.

- Sequential stages. A 60% chance of passing stage one and a 50% chance of passing stage two given stage one produce $60\% \times 50\% = 30\%$ for both. Adding them would be wrong.

An assistant can help identify which structure is intended before anyone adds or multiplies, and can check that relevant alternatives have not been left out.

24. What other technical checks are useful?

Possible mistake	What an assistant can help check
Reversing conditional probabilities	The probability of the evidence given H is not the probability of H given the evidence.
Multiplying probabilities as if they were odds	Convert to odds, multiply by the likelihood ratio, then convert back.
Shifting the hypothesis mid-argument	Population, time period, threshold, and meaning stay the same throughout.
Ignoring how evidence was selected	How reports were found, published, remembered, or missed.
Overlapping or missing explanations	Overlaps are handled and plausible alternatives are included.
Judging a complex theory by its best version	Its predictions are averaged over its internal variants with justified weights.
Fragile conclusions	Which changes in priors, weights, or dependence would reverse the result.
False precision in the output	Exact arithmetic is distinguished from uncertainty in the inputs and structure.

25. Does AI make technical skill unnecessary?

AI can reduce the formal technique a scholar must personally execute, but reviewing definitions, assumptions and model structure remains essential. An assistant may miss dependencies or produce a convincing error. Agreement with a second AI is a weak safeguard when both share sources, training patterns or proposed inputs. Use ordinary calculation tools to check arithmetic, primary sources to check claims, and independent specialist review where warranted. No AI assurance substitutes for these checks.

Try it yourself: a starter prompt

"I want to estimate how plausible [claim] is. My prior is about [X]% because [reasons]. Here are the pieces of evidence I consider relevant, with my rough sense of their strength: [list]. Please (1) convert everything to odds and show each step, (2) flag any evidence that may overlap or be counted twice, (3) tell me if any alternative explanation is missing, and (4) show how the result changes if my prior or my strongest piece of evidence is set higher

or lower. Keep my inputs fixed and list your suggestions separately.”

Part 6. Fair objections, fairly answered

26. Isn't this just a complicated way to describe common sense?

It formalizes the best part of common sense: start from what is already known, examine the new evidence, and change your mind by the right amount. But arithmetic catches mistakes intuition misses, including neglecting base rates, overrating weak evidence, and counting one source several times. The benefit is consistency and inspectability, with the real possibility of better accuracy.

27. Is it useful when there are no repeated experiments?

Yes, as a framework for comparing uncertain explanations, which is much of what historians and many social scientists do. Without repeated observations the inputs are harder to justify, so a careful analysis separates measured frequencies from estimated probabilities and shows how sensitive the result is to disputed judgments. A calculated posterior is conditional on the model and its inputs; it is not an experimentally established frequency.

28. Aren't these numbers just subjective?

Some are judgments, some are estimated from data, and many combine both. Judgment is not the same as arbitrary choice: it can be constrained by evidence, comparison cases, mechanisms, and consistency checks. The fair objection is “why is this value or range justified?” The insufficient objection is “a judgment was involved, so no numerical analysis is legitimate.” Selecting sources, trusting testimony, and declaring an explanation convincing are also judgments. Transparency does not make a judgment correct, but it makes its correctness something we can discuss.

29. Can a Bayesian analysis be rigged to produce any answer?

A badly conducted one can. So can an essay, a choice of sources, or a regression. The safeguards are documenting inputs, testing alternatives, keeping background evidence separate from newly scored evidence, checking the model, and reporting results that cut against the author's preferred conclusion. Arithmetic does not cure motivated reasoning. It makes motivated reasoning easier to catch.

30. Doesn't enough evidence make the prior irrelevant?

Often the prior's influence shrinks as evidence accumulates, but there is no universal guarantee. Sparse data, ambiguous observations, poorly distinguished explanations, or a flawed model can leave assumptions very influential. More observations from the same biased process do not repair the bias.

31. Is everyone secretly doing Bayesian statistics?

No, and the argument here does not claim that. Frequentist methods estimate parameters and test hypotheses without assigning prior probabilities to them, and qualitative scholarship establishes facts and explores mechanisms without producing posteriors. Those are genuine alternatives. The narrower claim is this: whenever someone concludes how plausible an explanation is after weighing background knowledge and evidence, that conclusion rests on a starting assessment and on weights given to the evidence. Calling the account verbal does not exempt it from scrutiny.

32. Is Bayesian reasoning automatically more accurate?

No. It improves consistency, exposes assumptions, and makes forecasts testable, but it can also give a weak model an undeserved air of authority. Accuracy depends on the data, the alternatives considered, and the quality of the inputs. Good practice checks predictions and revises models that fail. Gelman and Shalizi argue that applied Bayesian work requires exactly this cycle of model checking, not just updating among a fixed set of hypotheses [1]. Their point strengthens the case for scrutiny; it does not weaken the case for explicit updating.

33. What if the hypothesis itself is vague?

Define it before assigning a probability. “The policy works” could mean any positive effect, a large effect, a cost-effective effect, or an effect in one population. These are different claims, and a precise number attached to an ambiguous claim is not a precise answer. Specify the population, outcome, time period, threshold, and alternatives. In historical work, distinguish “this report contains an accurate observation” from “the author’s explanation of it is correct.”

34. Does a high posterior prove causation?

No. Causal claims require assumptions about what would change under an intervention. Confounding, selection, and measurement error remain problems in Bayesian analysis, and a posterior built on an inadequate causal model can be confidently wrong.

35. Does a probability tell us what to do?

Not by itself. Decisions also depend on costs, benefits, and the consequences of being wrong. A 27% chance that a program helps may justify a cheap pilot study while falling far short of what a costly national rollout requires. The same probability can support different decisions when the stakes differ.

36. How should uncertainty about the inputs be reported?

A posterior expresses uncertainty within a chosen model. Varying disputed priors and likelihoods shows uncertainty about the model’s inputs. Neither captures omitted explanations or flawed measurement. For readers, give a rounded central result, a justified range, and the assumptions most capable of changing it, and do not label an arbitrary low-to-high spread a statistical interval.

Why are historical likelihoods often harder than the prior?

An absence of manuscripts is informative only relative to their expected production, preservation, discovery and recognition under each explanation. A matching phrase is informative only relative to its prevalence and the search that found it. These conditional predictions, often more than the starting prior, determine the evidential update. Judgmental ratios can be useful, but should be justified, bounded and distinguished from measured frequencies.

How can small errors compound?

If twenty evidence factors are each set to 2 when their justified value is 1, their product is 1,048,576. Even multiplying a factor of 1.2 twenty times gives about 38.3. These are illustrations of repeated directional error, not estimates of typical research bias. Grouping dependent observations is necessary but does not remove a shared tendency to exaggerate every group. Test several disputed inputs together and show how the conclusion changes.

Why does searching for evidence change its weight?

A phrase specified before examining a corpus differs from the best resemblance discovered after thousands of comparisons. The evidence includes the discovery process: the texts searched, candidate selection, spelling choices, discarded matches and reporting rules. A fair control repeats that process for other writers. Adding a modest discount to a selected match does not by itself calibrate its evidential value.

What if the alternatives are open ended?

For a catch-all hypothesis such as "someone other than Stratford," the likelihood is an average of the evidence probabilities under its component alternatives, weighted by their probabilities within that group before the update. Use neither the most favorable rival nor an imaginary person combining all rivals' strengths. A pooled residual cannot fairly establish individual ceilings without justified internal weights. Include an unknown alternative and test the effects of how it is modeled.

What is a p-value unable to tell us?

A p-value is not the probability that a candidate is the author, and missing a conventional threshold does not establish zero evidence. Nor can a positive-looking exploratory result be converted directly into an authorship probability. Examine the comparison population, selection process, uncertainty, power, and independent replication before assigning substantial evidential weight.

How should joint sensitivity be reported?

Vary disputed priors, likelihoods, dependence assumptions and residual allocations together, explaining the rationale for each combination. One-factor changes can miss a common directional bias. A neutral factor of 1 can withhold an unsupported assumed bonus in a diagnostic scenario; it does not establish that the underlying observation is literally valueless. If no model is adequately justified, label the calculations as scenarios and withhold a preferred posterior.

A compact checklist

Define the claim. Explain the prior. Compare how well each explanation predicts the observations. Account for selection and dependence. Update. Vary the disputed inputs. Check the model against what actually happened. State what remains unresolved.

This sequence makes an analysis open to correction rather than merely open to admiration.

Part 7. A friendly challenge to the skeptic

37. What should we say to “You cannot put a number on that”?

Ask what supports the proposed numerical estimate. A critic may reject a value, dispute the model, or decline to quantify. A broad range is useful only if its boundaries have a reasoned basis. The legitimacy of another person's verbal assessment does not supply the missing justification for our number.

If the alternative is “unlikely,” how unlikely? If it is “strong evidence,” how much more expected is the observation under one explanation than the other? If it is “the consensus,” what produced that consensus, and has its underlying evidence already been counted? These questions remain whether the answers are written as numbers or as adjectives.

A critic may reject a model or suspend judgment. What a critic cannot show, merely by refusing numbers, is that a confident verbal verdict rests on fewer assumptions. If the inputs cannot support a narrow percentage, widen the range. If they cannot support a probability at all, the verbal verdict should admit the same limitation.

38. What is the closing argument?

The positive case for explicit analysis is that it makes assumptions inspectable, calculations reproducible and disagreements testable. Those benefits support the method; they do not establish any particular likelihood ratio, starting probability or posterior. Qualitative and quantitative arguments should both explain how their evidence supports their conclusions.

Serious objections include poor reference classes, missing alternatives, unsupported likelihoods, double counting, search selection and models contradicted by observations. Answer them directly by revising the model, widening defensible bounds or withholding a numerical verdict. Pointing out that verbal arguments can also fail does not repair a defective quantitative analysis.

The useful questions are: what justifies each input; what information has already been counted; how was the evidence selected; which alternatives are missing; and what happens when several contested judgments change together? Explicit calculations help only when these substantive questions remain visible.

A discipline for explicit reasoning

An undisclosed assumption is not an absent assumption. An unquantified judgment is not a judgment-free conclusion.

AI makes explicit analysis easier to construct. That is an opportunity for better scrutiny and a risk of more polished overconfidence. The appropriate aim is a reviewable argument whose limits are as visible as its conclusions.

Appendix. Formulas, sources, and attribution

Converting between probability and odds. For probability p , $\text{odds} = p \div (1 - p)$. For odds of a to b , $\text{probability} = a \div (a + b)$.

One-step update. For prior probability p and combined likelihood ratio L , the posterior probability is $pL \div (1 - p + pL)$. For two hypotheses, $L = P(E \text{ given } H) \div P(E \text{ given the alternative})$. All probabilities are conditional on the same background information.

Sequential updating. After evidence $E1$ and $E2$: $\text{posterior odds} = \text{prior odds} \times [P(E1 \text{ given } H) \div P(E1 \text{ given the alternative})] \times [P(E2 \text{ given } H \text{ and } E1) \div P(E2 \text{ given the alternative and } E1)]$. Making the second factor conditional on $E1$ avoids assuming the observations are independent when they are not.

Checks. With combined factor 3.4, priors of 0.20, 1/3, and 0.50 yield about 0.459, 0.630, and 0.773. With prior odds 1 to 2, sequential factors 17, 0.5, and 0.4 give odds 3.4 to 2, or about 0.630. Sport factors 0.25, 0.4, and 0.6 yield about 0.515, 0.630, and 0.718. Percentages are rounded.

Sources

- [1] Gelman, Andrew, and Cosma Rohilla Shalizi. 2013. "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology* 66 (1): 8-38. DOI: 10.1111/j.2044-8317.2011.02037.x.
- [2] Kent, Sherman. 1964. "Words of Estimative Probability." *Studies in Intelligence* 8 (4): 49-65.
- [3] Mosteller, Frederick, and Cleo Youtz. 1990. "Quantifying Probabilistic Expressions." *Statistical Science* 5 (1): 2-12 (main article; discussion through 34).
- [4] Mastrandrea, Michael D., et al. 2010. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. Intergovernmental Panel on Climate Change.
- [5] Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90 (430): 773-795. DOI: 10.1080/01621459.1995.10476572.
- [6] Gigerenzer, Gerd, and Ulrich Hoffrage. 1995. "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats." *Psychological Review* 102 (4): 684-704.
- [7] National Library of Medicine. "Color Vision Deficiency." *MedlinePlus Genetics*. medlineplus.gov/genetics/condition/color-vision-deficiency/

Development and attribution. Rod Malpert proposed this FAQ, its progression from introductory explanation to methodological objections, and its central argument: verbal assessments of plausibility and evidence do not escape the starting assumptions and weighting judgments that explicit Bayesian analysis exposes. He requested the side-by-side example, the emphasis on sensitivity, transparency, and double counting, and the focus on AI as a tool that lowers technical barriers while leaving the researcher in control. ChatGPT assisted with the original drafting, mathematical verification, and document production. Claude (Anthropic) assisted with the October 2026 revision, including reorganization, the expanded odds tutorial, and additional sources. The identification example, sensitivity calculations, and closing argument are original

illustrations, not quotations from or assessments of particular correspondents. Malpert directed the argument and retains responsibility for the final judgments.

About the Author

Rod Malpert holds a J.D. and M.A. from Cornell University. His research uses human-AI collaboration and explicit probabilistic reasoning to examine philosophical, historical, and public-policy questions. He practiced law in Silicon Valley. He was a managing partner on the executive committee of Fragomen, Del Rey, Bernsen and Loewy, an Am Law 100 international law firm.

Website and current papers: RodMalpert.com