Appendix A. Prior Probabilities and Concerns Regarding "Double-Counting"

In Section 3.3, we argue that evidence is never "double-counted" in logical Bayesianism. To further expound this argument, suppose we were to accidentally incorporate information into the analysis twice, in that some fact X gets included in both the background information and the evidence: $I = I_0\, X$ and $E = E'\, X$. Using the product rule, our posterior probability on hypothesis H_i equals:

$$P(H_i \,|E\, I) = P(H_i \,|E'\, X\, X\, I_0) = P(H_i \,|E'\, X\, I_0) = P(H_i \,|E'\, I)\ , \qquad \text{(A1)}$$

because "X and X" in the second term of (A1) is logically equivalent to "X." Therefore, all of our updating in this case comes from E', and X affects our probabilities only through the background information. Alternatively, suppose we express two pieces of evidence as $E_1 = X_a\, X_b$ and $E_2 = X_b\, X_c$. The rules of probability dictate that $P(H|E_1\, E_2) = P(H|X_a\ X_b\ X_c)$, because "$X_b$ and X_b" is logically equivalent to "X_b," so considering E_1 and E_2 instead of first decomposing the information into X_a, X_b, and X_c should make no difference to the conclusion.

Nevertheless, one might worry that in practice we could fail to live up to the mathematical ideal of logical Bayesianism when assessing priors on an inductively-inspired hypothesis, in a manner that does lead in effect to double-counting the evidence—this is a variant of Concern (a) regarding subjectively-biased priors discussed in Section 5.1. The idea here is that something we consider to be evidence might inadvertently also have been part of the background knowledge that informed a prior, thus its effect mistakenly boosts the probability on the hypothesis twice.

Our response to this worry is twofold. First, the risk of substantial double-counting is low. For most research that involves original data collection (especially fieldwork), keeping track of what we classify as evidence E and what we classify as background information I is a straightforward task—background information includes what we read in literature prior to our original research, and evidence includes what we discovered thereafter. Alternatively, if the research agenda entails going back through existing research to reevaluate established rival hypotheses and assess the current state of cumulative knowledge, then we should start from equal (ignorance) priors and incorporate evidence from these existing studies into our analysis piece by piece. In this situation, there is no relevant background information to inform priors, and hence there is no risk of accidental double-counting.

Second, our guidelines for assigning priors in Section 5.1 guard against bias arising from subconscious, accidental double-counting of some information X in both I and E. These recommendations include placing a penalty on the prior of an inductively-inspired hypothesis relative to established rivals, using equal priors, assessing how sensitive the findings are to different priors, or simply focusing on assessing likelihood ratios for the evidence and allowing readers to use their own priors, drawing on their own background information. If readers find a hypothesis much less plausible a priori than the author, they will demand evidence that weighs more strongly in favor of that hypothesis before they concede that it provides a better explanation than the rivals. This dynamic occurs frequently during the peer review process, albeit informally.

Similar recommendations for handling priors are common in literature on Bayesian statistics. For example, Berger and Berry (1988:162) advocate focusing on likelihood ratios: "The investigator... need not be concerned with the initial probability [prior] chosen by a possible consumer of the data; it suffices for the investigator to show how the data will change this initial probability into a final probability [posterior]," while Greenland (2006:766) notes

that: "Acceptability of an analysis is often enhanced by presenting results from different priors..."

A further point on priors merits emphasis. Priors are sometimes held up as the most important feature of Bayesianism,[1] but this characterization is not accurate. The most fundamental aspect of Bayesianism that distinguishes it from frequentism is the very definition of probability—as rational degree of belief, rather than a frequency-based characteristic of a population. The most important inferential step in Bayesian reasoning (especially for qualitative research) is not assessing priors, but rather assessing likelihood ratios, which tells us how we should update our prior odds.

Ideally, one would like to be in a situation where the evidence overwhelms the role of prior probabilities. In qualitative research, assignment of priors will usually be the analytical step that is most subject to arbitrariness, because we cannot hope to fully list and carefully consider all elements of our background information that influence our beliefs about the plausibility of hypotheses. While background information also matters for assessing likelihoods, it is easier to identify a few key elements of I that matter most for $P(E|H\,I)$. Moreover, assessing (or approximating) likelihoods, $P(E|H\,I)$, while challenging, is an inherently easier task than assessing priors, $P(H|I)$, because evidence is concrete, specific, and observable, whereas hypotheses are abstractions that cannot be directly observed.

Let us return to our state-building example (Sections 3.2 and 4) and assess to what extent we should worry about the possibility of erroneously using the same information both to assess a prior and to update that prior. Suppose Study S in the state-building literature includes fact $X=E_2$, and analytical findings from this study inform our priors on H_W and H_R—but we forget that Study S included E_2. We then come across E_2 in a different context; it leads us to devise H_{LRA}, at which point we must go back to our background information and reassess priors for the three hypotheses. With regard to prior odds on H_W vs. H_R, forgetting that E_2 was actually in the background information on existing state-building literature does not create any bias, because E_2 does not discriminate between these two hypotheses—this evidence is essentially equally likely under each. If we then move forward and incorporate E_2 into our analysis as evidence (still unaware that I included E_2), the odds on H_W vs. H_R remain unchanged, again because E_2 is equally likely under each. In practice then there is no "double-counting" with respect to inferences about H_W vs. H_R.

Turning to prior odds on H_{LRA} vs. the rivals, the fact that state-building literature mentioned E_2 but we have forgotten is also irrelevant, because that literature does not discuss H_{LRA}, so any aggregate analytical findings we do remember from that literature have no direct relevance for priors on H_{LRA}. Similarly, analytical findings from literature on other macro-political outcomes that treats labor-repressive agriculture as a salient causal variable have no direct relevance for priors on H_{LRA}, because this literature (by definition) does not consider H_{LRA}. Any bias on H_{LRA}'s prior would therefore have to arise by digging through literature that includes E_2, noticing that E_2 is relevant for the relative plausibility of H_{LRA} vs. the rivals, including E_2 in the background information, and then going on to treat E_2 as evidence to update a prior that was already informed by E_2. Here we either have an obvious and rather improbable accounting error in classifying information as background vs. evidence, or there is a problem of subconscious bias in that the author knows E_2 is not part of the background information that should inform the prior

[1] As MacKay (2003:347) observes: "There is a popular myth that states that Bayesian methods differ from orthodox statistical methods only by the inclusion of subjective priors, which are difficult to assign, and which usually don't make much difference to the conclusions."

odds on H_{LRA} vs. rivals, but is not able to objectively assess the prior odds without subjective contamination from knowing about E_2. This situation brings us back to Concern (a) in Section 5.1 and our guidelines for protecting against such subconscious bias. In Section 4, we proceeded by considering both a low prior on H_{LRA} relative to H_R, and equal prior odds.

Appendix B. Resolving the "New Problem of Old Evidence"

We noted in Section 3.3 that scholars from the psychological/subject school of Bayesianism often diverge from the logical Bayesian tenet that relative timing is irrelevant for inference. We mentioned the example of Jeffrey's (1983) "probability kinematics," which allows the order in which evidence is learned to affect posterior probabilities—thereby violating the fundamental product rule of probability. Another salient example in philosophy of science is the so-called "new problem of old evidence" (e.g., Glymour 1980, Earman 1992), which poses that even within a Bayesian framework, we cannot learn from old evidence.

Glymour (1980) argued that if probabilities are evaluated at a time when evidence E is already known, then $P(E|I)=1$, which in turn directly implies that $P(E|HI)=1$. Substituting into Bayes' rule, he finds:

$$P(H|EI) = P(H|I)\times P(E|HI)/P(E|I) = P(H|I)\times(1)/(1)\,, \tag{B1}$$

which yields $P(H|EI)=P(H|I)$, such that "old" evidence purportedly cannot alter our degree of belief in hypothesis H.

From a logical Bayesian perspective, the flaw in this reasoning lies in confusing temporal relationships with logical ones. If we wish to evaluate probabilities in the light of knowing evidence E, then E must explicitly appear in our notation as "conditioning information" alongside I to the right of the vertical bar. In essence, Glymour can only assert that $P(E|EI)=1$. His argument then collapses, because Bayes' rule accordingly yields:

$$P(H|E(EI)) = P(H|(EI))\times P(E|H(EI)) \,/\, P(E|(EI)) = P(H|EI), \tag{B2}$$

such that $P(H|EEI) = P(H|EI)$, which we already knew from the logical identity $EE=E$. In debunking Glymour's argument, astrophysicist Bill Jefferys (2007:7) notes: "...what Glymour has actually proved is the (well-known) fact that...quite sensibly...[we] cannot validly manipulate the Bayesian machinery to get additional information out of information that has already been used."

The crucial point is that when evaluating probabilities, the information upon which we condition our probabilities does not include whatever is in our heads at a particular moment in time. Instead, we condition on propositions located to the right of the vertical bar, which are explicitly specified and assumed to be true.

Appendix C. *Ad-Hoc* Hypotheses and Occam Factors

Section 3.4 introduced the logical Bayesian concept of Occam factors, which penalize hypotheses that over-fit the data. This appendix discusses Occam factors in more detail and provides two examples to show how they can arise in qualitative research.

To appreciate the importance of Occam factors, it is worth stressing that over-fitting can be a major problem within a frequentist framework that does not allow prior probabilities on

hypotheses or fixed parameters. When working with quantitative datasets, analytical models can be made arbitrarily complex with a multitude of adjustable parameters that end up fitting not just the signal of interest, but the noise as well. Detecting over-fitting can be particularly challenging in orthodox statistics, because adding extra parameters can always improve the likelihood of the data under the model.

Within logical Bayesianism, however, an *ad-hoc* hypothesis that is too closely tailored to fit the arbitrary details of the data incurs a low prior probability via Occam factors that arise automatically from correctly applying probability theory. These Occam factors keep us from favoring an overly complex hypothesis compared to a simpler hypothesis that adequately explains the data.

Recall that generally speaking, an *ad-hoc* hypothesis is properly regarded as one member of a family of hypotheses characterized by multiple parameters that take on different, but equally arbitrary values. To restate this point in slightly different terms, an *ad-hoc* hypothesis emerges from a model with multiple parameters that *a priori* could have taken on a large range of different values. As a model becomes more complex, its prior probability becomes spread out over a larger parameter space, and the posterior odds are reduced to the extent that this parameter space must be fine-tuned to fit the observed data. Similarly, whenever we include another parameter in the model and find that the range of values it must assume to account for the data is much narrower than the prior range of values deemed feasible given the background information alone, the model receives an Occam penalty.

Whether the posterior odds favor a more complex model relative to a simpler model depends on whether the complex model fits the data sufficiently better to overcome its Occam penalty. Compared to complex models, simpler models are generally ruled out more easily, because they are less able to explain a diversity of possible outcomes. On the other hand, Bayes' theorem rewards the simpler model for sticking its neck out and making less flexible predictions if those predictions come true. Bayesian analysis therefore helps find the signal without over-fitting the noise.

To see how Occam factors emerge from the mathematics of Bayesian probability, we reconsider the card-draw example presented in Section 3.4 (adapted from Jefferys 2003), where we draw the six of spades from a deck held by a stranger at a party. We are interested in comparing two hypotheses: H_R = *The card was arbitrarily selected from a randomly shuffled deck*, and an *ad-hoc* rival, $H_{6\spadesuit}$ = *The stranger is a professional magician with a trick deck that forces the six of spades*. The first step is to recognize that $H_{6\spadesuit}$ is one member of a family of 52 equally plausible related hypotheses, $H_M = H_M\, c_1$ *or* $H_M\, c_2$ *or ... or* $H_M\, c_{52}$, where $H_M\, c_k$ = *The magic trick forces card* c_k. In other words, we must compare H_R against H_M, a more complex model with a parameter c_k that can be adjusted to fit the data. We wish to calculate the posterior odds:

$$\frac{P(H_M|E\ I)}{P(H_R|E\ I)} = \frac{P(H_M|\ I)}{P(H_R|\ I)} \times \frac{P(E|H_M\ I)}{P(E|H_R\ I)} \tag{C1}$$

Expanding the numerator of the likelihood ratio (the right-most term in C1), we have:

$$\frac{P(E|H_M\ I)}{P(E|H_R\ I)} = \frac{\sum P(c_k|H_M\ I)\ P(E|H_M\ c_k\ I)}{P(E|H_R\ I)}, \tag{C2}$$

where we have used the law of total probability to introduce a sum over all 52 possible values of the card parameter *c*. In essence, we are averaging the likelihoods under each sub-hypothesis in the magic-trick family, weighted by the prior probability that the card parameter takes a particular value. When we plug in the 6 of spades for the evidence *E*, the sum in the numerator picks out that single value for the parameter *c*, because the likelihood of E=6♠ is zero for every sub-hypothesis except for that which forces the 6 of spades:

$$\frac{P(E|H_M\ I)}{P(E|H_R\ I)} = \frac{(\frac{1}{52}\times 0 + \frac{1}{52}\times 0 + \cdots + \mathbf{\frac{1}{52}}\times \mathbf{1} + \frac{1}{52}\times 0 + \cdots)}{(\mathbf{\frac{1}{52}})}. \qquad \text{(C3)}$$

In the denominator above, we have used the fact that the likelihood of E=6♠ under the random draw hypothesis is 1/52. Substituting (C3) into (C1), we can now rewrite the posterior odds ratio as the product of three factors:

$$\frac{P(H_M|E\ I)}{P(H_R|E\ I)} = \frac{P(H_M|\ I)}{P(H_R|\ I)} \times \frac{\left(\frac{1}{52}\right)}{1} \times \frac{1}{(\frac{1}{52})}. \qquad \text{(C4)}$$

These three factors on the right-hand side of (C4) are the model-level prior, the Occam penalty—a factor of 1/52 in the numerator, and the "fitted likelihood"—a factor of 1/52 in the denominator. The model-level prior remains to be assessed, using any salient background information about the chances that the stranger is a skilled magician as opposed to an ordinary partygoer with a randomly shuffled deck. The Occam penalty arises from the prior probability that a magic trick would favor the six of spades. The fitted likelihood, $P(E|H_M\,6\spadesuit\,I)\,/\,P(E|H_R\,I)$, assesses how surprising or expected our evidence is under $H_{6\spadesuit}$ relative to H_R once we have chosen the six of spades as the parameter value for the magician model.

In essence, the more complex model H_M receives an Occam penalty when the data obtained rules out all but one of the 52 parameter values that were plausible *a priori.* This Occam factor keeps us from favoring the *ad-hoc* six of spades hypothesis, which on its own makes the card we chose much more likely than the random-draw hypothesis. Note that in general, the Occam factor will not exactly cancel the fitted likelihood; that effect is a special feature of this example. It is also important to emphasize that the posterior odds could end up favoring the more complex model, if the fitted likelihood is good enough to overcome the Occam factor. Accordingly, logical Bayesianism does not always favor simplicity—it balances simplicity against explanatory power.

A second example illustrates how Occam factors can emerge in explicit Bayesian process tracing.[2] Suppose we have two plausible explanations for why the government of Gonduria, a developing country on the Pandor continent, expanded social programs to reach a larger proportion of the poor:

H_{WB} =*Expanding social programs was a condition for a World Bank loan;*

H_R =*The government designed these measures to improve its approval ratings after the latter dropped below a critical threshold,* r_c.

[2] For the sake of illustration, we are explicitly identifying and evaluating an Occam penalty, but Occam factors arise automatically if Bayesian analysis is correctly employed. In actual practice, we need not think about Occam factors as a separate step in Bayesian analysis.

H_R denotes a family of hypotheses, where r_c could take on many different values. A priori, it would be reasonable to assume that the threshold rating r_c falls between 25% and 50%. Regarding the upper limit, we reason that democratic governments tend to become concerned once approval ratings drop below 50%. We set the lower limit drawing on background information that approval ratings in Pandorian democracies generally have not dropped below 25% during periods of normal politics. We wish to calculate the posterior odds ratio (equation 3) for the two hypotheses in light of evidence E_0 =*The government's approval rating at the time, r*, was 44%.*

We begin by evaluating the likelihood of the evidence under H_R:

$$P(E_0|H_R\ I) = \sum P(r_c|H_R\ I)\times P(E_0|r_c\ H_R\ I) \tag{C5}$$

where as in the previous example, we have used the law of total probability to introduce a sum over all possible values of the critical threshold (recall that each value of r_c defines a specific hypothesis in the H_R family); for simplicity we sum over integers instead of integrating over a continuum.[3] When r_c >50% or <25%, we have $P(r_c|H_R\ I)$=0. We take the prior likelihood of the threshold parameter to be uniform over the range of 25%–50%, such that $P(r_c|H_R\ I)$=1/25. Denoting evidence E_0 as r^*=44%, we have:

$$P(E_0|H_R\ I) = (1/25) \sum P(r^*=44\% \mid 25\% \leq r_c \leq 50\%\ H_R\ I) \tag{C6}$$

The summand vanishes unless $r_c \geq$44%; otherwise the threshold hypothesis would be contradicted. For $r_c \geq$44%, we take all values of $P(r^*=44\% \mid r_c\ H_R\ I)$ to be equal, assuming that approval ratings at the time the government expanded social spending are independent of the critical threshold.[4] We can then replace the sum in equation (C6) with a factor of 7:

$$P(E_0|H_R\ I) = (7/25)\ P(r^*=44\% \mid 44\% \leq r_c \leq 50\%\ H_R\ I) \tag{C7}$$

More generally, for evidence E that includes r^*=44% along with other salient observations, we have:

$$P(E|H_R\ I) = (7/25)\times P(E \mid 44\% \leq r_c \leq 50\%\ H_R\ I) \tag{C8}$$

We can now calculate the posterior odds ratio for H_R vs. H_{WB}:

$$\frac{P(H_R|E\ I)}{P(H_{WB}|E\ I)} = \frac{(7/25)\times P(H_R|I)\times P(E \mid 44\% \leq r_c \leq 50\%\ H_R\ I)}{P(H_{WB}|I)\times P(E|H_{WB}\ I)} \tag{C9}$$

We find that H_R is penalized relative to H_{WB} by an Occam factor of 7/25, regardless of how plausible we find the family of hypotheses H_R relative to the World Bank hypothesis. This moderate penalty arises because the data r^*=44% rules out a moderate portion of the parameter space judged feasible given the background information. Had the value of r^* been lower, the Occam penalty would have been less significant. If the government's approval ratings at the time fell below 25%, this evidence would be consistent with any value of the threshold between 25–50%, and H_R would not incur an Occam penalty relative to H_{WB}.

[3] We would not expect arbitrarily close values to be observationally distinguishable so this approximation seems reasonable.

[4] This assumption is an oversimplification—there could be many dependencies.

References

Berger, James, and Donald Berry. 1988. "Statistical Analysis and the Illusion of Objectivity," *American Scientist* (March-April):159-165.

Earman, John. 1992. *Bayes or Bust?* Cambridge: MIT Press.

Glymour, C. 1980. *Theory and Evidence.* Princeton University Press.

Greenland, Sander. 2006. "Bayesian Perspectives for Epidemiological Research," *International Journal of Epidemiology* (35):765-775.

Jefferys, William. 2003. Book review: "Bayes' Theorem," *Journal of Scientific Exploration* 17(3:537-42).

_____. 2007. "Bayesians *Can* Learn from Old Data." https://repositories.lib.utexas.edu/bitstream/handle/2152/29425/BayesiansOldData.pdf?sequence=1

MacKay, David. 2003. *Information Theory, Inference, and Linear Algorithms.* Cambridge.