Title page

Article title: The importance of prior choice and specification in practical analyses has been over-emphasised

Authors: William D. Pearse[1*]

[1] Department of Life Sciences, Imperial College London, Ascot SL5 7PY, United Kingdom. ORCID: 0000-0002-6241-3164.

[*] To whom correspondence should be addressed: will.pearse@imperial.ac.uk.

Acknowledgments: I am grateful to XXX anonymous reviewers, and the editorial board, for their help improving this manuscript. XXX provided insightful comments on the manuscript. The Pearse lab and I are funded by XXX.

1 Abstract

Bayesian statistical methods have become commonplace in the life sciences, and differ from frequentist methods in many ways including their requirement of the specification of *a priori* information ('priors'). Each coefficient in a Bayesian model requires a defined prior distribution, and the selection and parameterisation of prior distributions affects the results generated from a Bayesian model. Since priors reflect literal beliefs and are challenging, if not impossible, to empirically justify, priors have been heavily studied and are a major cause of concern for those learning Bayesian methods. While the defining role priors play in Bayesian models is irrefutable and clear to all, here I argue that their biological significance in practical model-fitting has been over-exaggerated. I begin by showing how little they contribute to a posterior estimate by walking through the calculation of posterior samples, and then explore how badly specified a prior would have to be to impact a study with reference to toy examples and classic papers. I define the kinds of priors to beware—"anti-Cromwellian priors"—and give a practical rule of thumb—more than ten data-points per coefficient—to follow in Bayesian model-fitting. I then outline steps that can be taken in model fitting and criticism that would, I argue, more practically improve the fit of models while also detecting any impacts of priors were they to be present.

2 Main text

The last decade has seen an explosive increase in the use of Bayesian statistical methods (Gelman et al., 2014), likely in response to the development of new computational tools that make it easier than ever to fit models (Carpenter et al., 2017). I see this as a good thing because Bayesian methods are so (1) usefully flexible, (2) intuitive to interpret, and (3) straightforward to teach. (1) While software exists to make it easy to perform (standard) Bayesian regressions, most software is flexibile enough to easily specify and fit complex model structure that well-describe biological reality and its diverse forms of uncertainty (**Dietz2017**). (2) Bayesian results can be described in intuitive ways that make sense to non-scientists because they permit estimates of model probabilities. For example, I suggest that statements such as "90% more likely that there is a positive relationship than not", which is impossible to generate under a frequentist framework ("a less than 5% probability of seeing a slope this extreme were there no relationship" does not mean the same thing as the previous statement). (3) Because I, like many of my colleagues, were first trained in frequentist methods, it is easy to forget how counter-intuitive they are to students and how comparatively straightforward Bayesian methods are. Consider how long it takes to teach an undergraduate class what 'rejecting H_0' means and that the American Statistical Society were forced to establish a task force because of "the frequent misinterpretation of p-values".

There is, however, one aspect of practical Bayesian analysis that I think has been over-emphasised and mis-represented in the life sciences: the importance of prior distributions. Here I argue that the current magnitude of emphasis on priors, both when teaching and conducting practical analysis, is unnecessary, and that our time would be better spent carefully exploring models using methods such as posterior predictive checks. I can speak only to the life sciences, and perhaps ecology, evolution, and conservation, as they are my main area of expertise, but I think it is likely that my argument applies more broadly as well. I wish to state clearly that I am in no way claiming or arguing here that priors do not matter: their central and defining role in Bayesian model specification is irrefutable. It is of course (as I shall demonstrate) possible to torture any combination of model and dataset until priors are the determinant. I am, however, arguing that *the practical importance of prior specification is vastly over-exaggerated*. Everything matters: it is the role of statistics to

shine a light on what matters most.

A brief introduction to practical Bayesian model-fitting and priors

Briefly, the goal of a practical Bayesian analysis is to generate a posterior distribution of a model that well-describes data and can be used to derive insight about those data. Better and more complete descriptions of Bayesian statistics are given elsewhere (**Gelman2013**), although below I outline the points sufficient to follow my argument. The first step in this is to describe a model; for instance, if we were attempting to model the airspeed velocity of a sparrow as a function of how laden (weighed-down) it is, we might fit a linear regression model of the form:

$$\mu = a + bx \tag{1}$$

Where μ is the predicted, continuous speed, a is an intercept, and b a continuous slope for the effect of additional mass (x; an 'explanatory' continuous variable). This, in turn, would be mapped onto the response variable (y; the velocity) with an estimated amount of variation (variance; σ^2) as follows:

$$y \sim normal(\mu, \sigma^2) \tag{2}$$

. In a standard frequentist model, we would likely not directly estimate σ^2 but rather measure it indirectly through some measure like r^2, but this model is essentially unchanged across the two statistical schools of thought. The difference, in a practical sense, is how the model is fit to data, and whereas a frequentist model would generate 'point estimates' of each coefficient with associated estimates of uncertainty. These terms would be found by numerically maximising the joint likelihood of the data given the specified model, or through some analytical solution that has been found from those assumptions (Edwards, 1984). A Bayesian model seeks to find a posterior distribution for each coefficient (a, b, and so their composite μ, as well as σ^2) that can be summarised (*e.g.*, the median) to generate insight (Gelman et al., 2014). This posterior distribution of the model's coefficients

Velocity	Weight
2	4
3	3
4	2
5	1
6	0

Table 1: **Data used for example calculation of posterior estimate**. Please imagine the data repeated, with each entry appearing 6 times, making a total of 30 estimates.

is found using a numerical approximation method such as Markov Chain Monte Carlo (MCMC), ultimately relying on the following equation:

$$posterior = likelihood \times prior \tag{3}$$

where *posterior* is the posterior distribution of the parameters of interest (*e.g.*, b), *likelihood* is the same likelihood used in maximum likelihood methods (there is sometimes some disagreement as to terminology, but the equations used are the same), and the *prior* is an *a priori* statistical statement about the relative likelihood (see caveat to this terminology above) of all model coefficients, specified with a distribution. In passing, I note that frequentist methods also require the use of an *a priori*, axiomatic belief: that the long-term frequency of events, in the limiting case, is equal to their true probability. This belief has been tested, notably by tossing a coin thousands of times (Kerrich, 1950), but prior belief in Bayesian methods remains controversial and the subject of many papers (Banner et al., 2020). It is the goal of this essay to establish the follow rule of thumb: *if sufficient data collection has been carried out, prior definitions rarely matter*. I suggest that this could be combined with that other classic rule of thumb in statistics to '*have at least ten samples for each parameter to be estimated*', to make a joint rule of "*ten data-points per prior*". Like all rules of thumb, it is not always right (no rule can be simultaneously always applicable and always right; Gödel, 1931), but I suggest it is helpful to bear in mind.

A worked example of the role a prior plays in model-fitting

I feel the importance of a prior is best demonstrated by walking through the calculation of a single MCMC iteration's posterior probability. What follows is a walk-through for one such iteration (step) in the hypothetical model above. Imagine we have given the computer 30 estimates of bird additional weight (given in table 1), and are on iteration 1500 of a 2000 iteration process. Imagine also that the true model that generates the data in table 1 is a very simple equation—$\mu = 6 - 1 \times x = x$—the intercept is 6 and the slope is -1.

We have three coefficients in our model—a, b, and σ—and, as such, must specify at least three priors (one for each). Sensible priors might be the following:

$$a, b, \sigma \sim normal(0, 1) \tag{4}$$

The only prior we will focus on is the slope coefficient (b), and since my purpose is to show how priors don't matter, let's set a ridiculous one for the slope:

$$b \sim normal(5, 1) \tag{5}$$

Let us just sit, for a moment, with how absurd this definition of a slope prior is. We are saying that, prior to conducting the experiment, we are absolutely convinced (as represented by a variance term on the prior of 1) that birds fly faster when they are carrying more (the slope is positive and has a large value relative to the data). Of course these priors should probably been determined before the experiment, but still, we move on.

When calculating the posterior probability for this particular iteration, we need to know the coefficient terms in that iteration: let us say that $a = 6.1$, $b = -1.0$, and $\sigma = 1$. Our probability will always be $posterior = likelihood \times prior$, and I will estimate the posterior under sensible and absurd priors. I am using non-standard, abbreviated notation to save space: $p(d|m)$ where I will estimate the probability of seeing the $d$ata (the observed speed) given the $m$odel prediction. I will write each out in full, and will highlight in red font the bit that changes in the equation between

the two:

$$
\begin{aligned}
\left(liklihood\right) \times \left(prior_{sensible}\right) = \\
&\Big(p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05) \times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05) \times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05) \times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05) \times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05) \times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\Big) \times \\
&\Big(p(a = 6.1) \times p(b = -1) \times p(\sigma = 1)\Big) \\
&= \\
&\Big(0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398 \times \\
&0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398 \times \\
&0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398 \times \\
&0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398 \times \\
&0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398 \times \\
&0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398\Big) \times \\
&\Big(0.000000000331 \times 0.242 \times 0.242\Big) \\
&= 0.000000000000000000000164
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
\Big(liklihood\Big) \times \Big(prior_{sensible}\Big) = \\
&\Big(p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05) \times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05) \times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05) \times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05) \times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05) \times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\Big) \times \\
&\Big(p(a = 6.1) \times p(b = -1) \times p(\sigma = 1)\Big) \\
&= \\
&\Big(0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398 \times \\
&0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398 \times \\
&0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398 \times \\
&0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398 \times \\
&0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398 \times \\
&0.391 \times 0.394 \times 0.397 \times 0.398 \times 0.398\Big) \times \\
&\Big(0.000000000331 \times 0.00000000608 \times 0.242\Big) \\
&= 0.00000000000000000000000000000412
\end{aligned}
\tag{7}
$$

Blink and you will miss the difference. Despite the fact that the equation for the posterior distribution ($posterior = likelihood \times prior$) makes it looks as though prior and likelihood have equal weight, there are thirty terms for the data in each of the above calculations and only three for the priors. Every data point adds another term to the posterior probability formula while each prior

can only ever have one term. Thus, for example, if we had collected 100 pieces of data but kept every other part of the model the same, then using the same replication of data as outlined in table 1 we would have an equation as follows (I am not going to work this through to its numerical conclusion as I sense my point has been made):

$$
\begin{aligned}
\left(liklihood\right) \times \left(prior_{sensible}\right) = \\
&\Big(p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\times \\
&p(2.1|2.2) \times p(3.05|3.15) \times p(4|4.1) \times p(4.95|5.05) \times p(5.95|6.05)\Big)\times \\
&\Big(p(a = 6.1) \times {\color{red} p(b = -1)} \times p(\sigma = 1)\Big)
\end{aligned}
\tag{8}
$$

I emphasise that these calculations are usually performed on logarithms because the numbers become so small so quickly, such that even the modest differences above are quite quickly lost. Indeed, even quite large differences rarely functionally affect estimates of coefficients from posteriors (as we will see below in simulations), since it is the *relative differences* in posterior probabilities that are important. Thus an absurd prior ceases to matter quite quickly, because all coefficients are essentially equally unlikely (although see below a caveat for broad priors). Our absurd prior is having no impact whatsoever on the *relative* probabilities of our slope estimate because it's essentially constant around the true slope: for the absurd prior $p(b = -1.2) = 1.79 \times 10^{-9}$, $p(b = -1) = 6.08 \times 10^{-9}$, and $p(b = -0.8) = 1.98 \times 10^{-9}$. Sure, these values are different, but in the context of the other 100 data-driven terms in the model they are negligible. Statisticians often use the phrase 'swamping the prior' or 'swamping the data': in almost every conceivably sensibly chosen model-fitting exercise, the number of data terms is so large that it 'swamps' the prior and renders it functionally unimportant.

But what about non-conjugate priors and Cromwell's Rule?

The sceptical, statistically-informed reader is perhaps now thinking of non-conjugate priors and "Cromwell's Rule". Indeed, I very much hope so, because they are the exceptions that prove my rule of thumb.

Informally, a conjugate prior is a prior distribution that is mathematically convenient for whatever distribution its coefficient sits within. For example, in equation 4 the distributions for a and b (normal distributions) are appropriate for the mean (scale) of a normal distribution, whereas the distribution for σ is not (you cannot have a negative variance and the normal distribution has no finite limits). There is a great deal of very important statistical work that goes into determining the appropriate conjugate prior for each parameter in each distribution, but the empirical life scientist does not spend their time looking these up, but rather simply fits whatever is appropriate (and this is often the default option in a model).

I would also (perhaps controversially) propose that the practical impacts even of non-conjugate prior specifications on the posterior distribution are often negligible. Most MCMC software contains

'warm-up' algorithms that change the proposal distributions after an initial search (*e.g.*, Bouckaert et al., 2014; Carpenter et al., 2017), such that proposed moves in later iteration are more limited and don't pass into the range of impossible coefficients anyway. Thus, in my own experience, even quite plainly absurd prior distributions fit just fine (perhaps with some flagged impossible moves in the warm-up phase). The distribution may have an impact on the speed with which warm-up is attained, but there is no aspect of the discussion that follows that doesn't apply to non-conjugate priors just as easily. They are, of course, improper, and they do, of course, slow model convergence, but in practical cases with sufficient data there is no reason to suppose they will have a large impact.

Cromwell's Rule, on the other hand, is a very real problem. The rule is useful but the name is unfortunate because of the genocidal campaign of Cromwell, although its origins are in his writing "*I beseech you, in the bowels of Christ, think it possible that you may be mistaken*". The essential idea is that you should never, ever specify a prior that makes any particular model coefficient impossible (*i.e.*, have a likelihood or probability of 0). Thus the following prior for b would be wrong because it makes it impossible for our slope to be positive at all.

$$b \sim Uniform(-100, 0) \tag{9}$$

Such priors are, in essence, the exception that proves the rule because they emphasise the importance of not picking priors that make any particular outcome essentially impossible. I might argue that equation 5 is essentially an impossible prior because of how narrow and how absurd the central value is. I will refer to such priors below as 'practical Cromwellian' priors, since they are not quite Cromwellian but have essentially the same impact as them.

Thus, to summarise, I yield that the choice of prior should not be defined so as to make possible things impossible (Cromwell's Rule), or impossible things possible (non-conjugate priors).

But what about cases where you have limited data?

In the event that you have essentially no data, then a prior obviously determines the answer you get from your model and so of course the prior is important. We have lots of fancy terms for this (*e.g.* 'swamping the data', as has already been mentioned), but I wonder sometimes if these terms are used to obscure an uncomfortable truth: if your prior is so important in determining your answer, you are explicitly stating that either you need more data or that you were so certain to begin with that you needn't have bothered starting the analysis in the first place. I note that much of the (I stress, excellent and needed) statistical work on prior definitions, and much of the advice about priors that is given in good textbooks, is focused on cases where data are limited.

To make this more explicit, figure **??** shows what happens when I simulated data of the kind outlined above. Specifically, I simulated an explanatory variable ($x \sim normal(0, 1)$) with a given number of draws ($n = 10, 20, 30, ...200$), and then a matching continuous response variable ($y \sim normal(x, 1)$). Using `rstanarm::stan_glm` I used auto-scaling, default priors on all coefficients other than the slope, which was given a normal distribution with a variance of 1 and a varying scale ($\mu = -1, -5, -10, ... - 60$). Repeating this exercise across all possible combinations of the number of draws and the scale parameter on the prior, with over 20 draws (datapoints) only a scale of XXX was sufficient to affect the estimate of the slope. Indeed, I invented the term 'practical Cromwellian' to cover the cases where I had 200 draws, since a slope of XXX was required to have any practical effect.

There are, of course, fields of study where priors have large impacts. In my own phylogenetics work, I am only too aware of the role that priors can play in analysis. But, telling, this is almost always touted as a problem to be solved because it reflects the fact that we have insufficient data and our models make insufficiently strong predictions about them. Thus I do not see such cases as something to celebrate, but rather to drive the development of new methods and approaches.

A popular example of the importance of prior specification is Link2013, who are widely cited as showing that a Uniform prior biases estimates of abundance (see below also for too broad priors). Yet this example is, if anything, a demonstration that the underlying concern is not prior specification

but rather limited data. Link2013 is a statistical report, and so arguably is not addressing a broad empirical problem in ecology but rather helping develop methods and clarify statistical philosophy in a problem of interest to statistical ecologists. I doubt Link2013 would disagree too strongly with me that this was their audience; the article is dominated by equations (I count 14 lines of pull-out equations and a two-page BUGS code box) and it focuses on a classic dataset that they themselves cite as being "*analyzed compulsively by statisticians*" Royle2007. It is so-cited and studied because it is a difficult dataset and difficult datasets are great for testing problems: the problem here is *what to do with such small datasets that you must artificially inflate them.* The empirical demonstration of the impact of priors is on six surveys—$n = 6$—where 68 individuals were (sometimes repeatedly) captured and each has a parameter of observability associated with it. Thus these data are well beyond any standard rule of thumb for estimating models (making them perfect for such a statistical exploration), but not ideal for demonstrating the kinds of real-world issues ecologists might face. Finally, however, when we look at the impact of changing the priors on the median estimates from the posterior distributions, we find that the two biased (wrong) priors estimate 95 and 98 hares, but the best prior estimates 93. If one of the textbook, classic examples of the impact of prior mis-specification is a difference of 5 individuals (all estimates have large 80% credibility intervals broadly in line with each other), I question whether the biological significance of this worst-case example is being somewhat overplayed.

But what about too-broad (too-weak) priors?

This is, in a sense, a special case of the above examples, but one that has received a lot of attention because it is so counter-intuitive. The argument (which is valid), is that overly broad—essentially too uncertain—priors can cause bias in models. This is excellently reviewed in Banner et al. (2020), but it normally results from a non-linear change of scale somewhere in the model such that a very broad, very uncertain prior distribution's long tails get unevenly compressed into a smaller region of space in another part of the model.

This problem serves as an excellent case study of what is a common pattern in the discussion of priors in the literature: the problem is outlined, it is then verified, but then there is no magnitude

as to the relative importance of the issue. Banner2020 is the best review article on Bayesian priors I have read, but it too is an example of this issue: figure 1 gives equal space and attention to the prior and the likelihood (despite, as we have seen above, the fact that they are uneven) and (correctly) plots the impact of a mis-specified prior in figure 2 without highlighting that what is plotted is the prior and not the impact of the prior on a model itself.

To make this even clearer, in figure **??** I simulate data essentially identical to the previous simulation, but this time changing the variance of the prior and using a logistic regression to match the example given in Banner2020. Specifically, I simulated an explanatory variable ($x \sim normal(0, 1)$) with a given number of draws ($n = 10, 20, 30, ...200$), and then a matching binary response variable ($y \sim Binomial(XXX)$). Using `rstanarm::stan_glm` I used auto-scaling, default priors on all coefficients other than the slope, which was given a normal distribution with μ of 0 and variances to match those highlighted in Banner2020 ($\sigma^2 = 2, 9, 100, 10,000$). In only the highest variance on the prior—which, for clarity, is a variance of *ten thousand* being fit to data that have a variance of 1, and so is four orders of magnitude higher than observed in the data—do we see any appreciable impact on the model outcome.

I do not disagree with the existing literature and happily restate the important point that what I have termed anti-Cromwellian priors are a source of bias and conservative priors do appear the better choice. But this study also highlights that this problem does not exist if there is sufficient data.

But what about the importance of prior work?

Priors are touted as a positive feature of Bayesian analysis, allowing us to incorporate prior work and explore the role uncertainty plays in our analysis. On the basis of the above, however, we can see that this may be theoretically true but is, in practical terms, at best a distraction. If setting a prior to be absolutely five times greater than a given slope in the opposite direction has essentially no impact on the outcome of a model with XXX measurements, then I do not see how any reasonably prior based on past experience could matter in an experiment. Indeed, I think arguing that biologists who have dedicated time, in many cases decades, to their systems should

view what are relatively minor adjustments to priors as important is an insult to their time. Most importantly, it massively down-plays the huge advantage of Bayesian approaches: their flexibility. Instead of tweaking a model's priors, to little apparent effect, why not make a better model that is more comprehensive and can include all of the data collected—including from prior experiments? Such arguments, ironically, used to be deployed in favour of using maximum likelihood methods over Bayesian ones, I note.

Conclusion: there are more things in heaven and earth than are highlighted in prior checking

As I outline above, if your model is so poorly specified that priors swamp it then this is obviously something to be concerned about. But, ultimately, time is not infinite, and there are many more productive and instructive things to consider first *before* checking the impact of priors. I frequently find quite advanced researchers who are fitting models and checking their prior distributions (and parameters) but have not done the following.

1. Checked chains for convergence and mixing (or even plotted their traces)
2. Performed posterior predictive checks (the simplest of which include just plotting the model predictions against the data) to see if the model is performing
3. Scaled and transformed their explanatory variables to make their coefficient comparable and ensure model convergence
4. Checked the r^2 of their models to see if they are performing adequately, or the relative pooling of their hierarchical terms to see what is driving variance

Of these, the first is a literal requirement of all Bayesian analysis, and the second strongly recommended. The third is also a requirement of many statistical methods for them to work, but also permits simple interpretation of model outputs. If you have ever wanted to know whether one factor has a greater association with a factor than another, then following step three will permit that. Step four is also, I would argue, the most important (the only?) steps in model criticism that matter and will, for the record, also highlight if you prior is mis-specified. By all means, re-fit

under different priors if you must. But please save yourself some time and check the things that are, *a priori*, most likely to matter: and they are not your prior definitions.

References

Banner, K. M., Irvine, K. M., & Rodhouse, T. J. (2020). The use of bayesian priors in ecology: The good, the bad and the not great. *Methods in Ecology and Evolution*, *11*(8), 882–889.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). Beast 2: A software platform for bayesian evolutionary analysis. *PLoS Computational Biolology*, *10*(4), e1003537.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Edwards, A. W. F. (1984). *Likelihood*. CUP Archive.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 3). Taylor & Francis.

Gödel, K. (1931). Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für mathematik und physik*, *38*(1), 173–198.

Kerrich, J. E. (1950). *An experimental introduction to the theory of probability*. Belgisk Import Company.