Diametros 17 (2020), 66: 10–33
doi: 10.33392/diam.1587

The Problem of New Evidence: P-Hacking and Pre-Analysis Plans

– Zoë Hitzig –
– Jacob Stegenga –

Abstract: We provide a novel articulation of the epistemic peril of p-hacking using three resources from philosophy: predictivism, Bayesian confirmation theory, and model selection theory. We defend a nuanced position on p-hacking: p-hacking is sometimes, but not always, epistemically pernicious. Our argument requires a novel understanding of Bayesianism, since a standard criticism of Bayesian confirmation theory is that it cannot represent the influence of biased methods. We then turn to pre-analysis plans, a methodological device used to mitigate p-hacking. Some say that pre-analysis plans are epistemically meritorious while others deny this, and in practice pre-analysis plans are often violated. We resolve this debate with a modest defence of pre-analysis plans. Further, we argue that pre-analysis plans can be epistemically relevant even if the plan is not strictly followed—and suggest that allowing for flexible pre-analysis plans may be the best available policy option.
Keywords: Bayesian confirmation theory; pre-analysis plans; replication crisis; predictivism; p-hacking.

Published online: 5 October 2020

§1 Introduction

'P-hacking,' a term used widely in contemporary scientific discourse, refers to a variety of practices. The phrase originates in the ubiquitous statistical method of reporting 'p-values' of statistical analyses of data. It has become the norm in scientific publishing to describe p-values below some threshold (often 0.05) as "statistically significant." Researchers may thus "hack" their analyses and reports of analyses in order to produce p-values that qualify as statistically significant. Since statistically significant results are

Zoë Hitzig
Department of Economics
Harvard University
1805 Littauer Center
Cambridge MA 02138, USA
email: zhitzig@g.harvard.edu
Jacob Stegenga
Department of History and Philosophy of Science
University of Cambridge
Free School Lane
Cambridge 3B2 3RH, UK
email: jms303@cam.ac.uk

more likely to be published and promoted, researchers have incentives to engage in such activities. P-hacking and the related issue of publication bias are frequently cited as twin evils at the core of the ongoing 'replication crisis' in the social and medical sciences, and are the targets of proposals for policy reforms intended to rework the incentive structures of scientific publishing.

What exactly is p-hacking? To illustrate the range of methods described as p-hacking, consider the following recent claims about p-hacking from researchers in a variety of disciplines. An influential definition from psychologists Simmons et al. (2011) is that p-hacking amounts to "flexibility in (a) choosing among dependent variables, (b) choosing sample size, (c) using covariates, and (d) reporting subsets of experimental conditions." Thomas Insel, a former director of the National Institute for Mental Health, summarizes the definition in Simmons et al. (2011) in slightly different terms: p-hacking, he holds, "refers to the practice of reanalyzing data in many different ways to yield a target result." Meanwhile, biologists Head et al. (2015) write, "'p-hacking' occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant" and when, the authors continue, "researchers try out several statistical analyses and/or data eligibility specifications and then selectively report those that produce significant results."

These claims illustrate the variety of notions of 'p-hacking' floating around in contemporary discourse. Further complicating matters, the term 'p-hacking' is often used interchangeably with other terms including 'data-dredging,' 'data-mining' and 'specification searching.' To summarize, the term encompasses practices including: performing multiple measurements in an experiment, arbitrarily trimming data, performing many post-hoc analyses of data, and selectively reporting only statistically significant results.

Many scientists claim that methods described as p-hacking are epistemically nefarious, because these practices are liable to lead to the detection of false-positive findings. However, these practices may also lead to the detection of true-positive findings. Some scientific disciplines *require* the use of methods that are, broadly construed, constitutive of p-hacking, such as astronomy and genetics. Such methods are used to discover true facts of nature, goes this view; to constrain the use of these methods would amount to constraining discovery.[1] Views on p-hacking have recently become vitriolic, in part due to the replication crisis in the social sciences and medicine. Some scientists consider p-hacking to be the very antithesis of science (Ioannidis, 2005, 2008; Simmons et al., 2011; Camerer et al., 2018), while other scientists consider methods that seem to constitute p-hacking to be just as scientific as any other legitimate mode of scientific inquiry (Pagan, 1987; Phillips, 1988; Backhouse and Morgan, 2000).

Though the epistemic issues that arise from such methods are often articulated in the context of frequentist statistics, in this article we investigate these issues through an appeal to Bayesian confirmation theory. We employ this formal framework to argue

[1] A distinction that might occur to readers is that between exploratory data analysis versus confirmatory data analysis. One might hold that the methods mentioned in this paragraph are merely exploratory, and thus ought not be deemed p-hacking. However, for reasons that follow later in this paper, little stock should be placed on this distinction. In short, our arguments that follow entail that analyses that are intended as confirmatory might not in fact add much confirmation yet might have some exploratory value, and analyses that are intended as merely exploratory might in fact end up offering a great deal of confirmation.

that in some cases, but not all, particular methods commonly referred to as p-hacking are indeed epistemically pernicious. We articulate the precise formal conditions under which this is so. To show that our conclusion is not sensitive to our choice of a Bayesian framework, we articulate a similar argument using model selection theory.

We then turn to a discussion of pre-analysis plans, which are a methodological device widely used to mitigate the peril of p-hacking.[2] Some say that following a pre-analysis plan is epistemically meritorious while others deny this: on the one hand, they might mitigate p-hacking and publication bias (and thus mitigate false positive findings), but on the other hand, if strictly followed, they constrain the freedom of scientists (and thus might mitigate true positive findings). In practice pre-analysis plans are often violated (Dwan et al., 2008; Casey et al., 2012).[3] We use the formal groundwork developed earlier in the paper to resolve the debate about the scientific legitimacy of such plans, offering a modest defence of the use of pre-analysis plans.

The epistemic peril of p-hacking is typically described in the context of frequentist statistics, in which analysts run multiple tests in the search for statistically significant 'p-values'. Roughly, the worry is that statistically significant findings can arise due to chance or bias in the data-generating method, so for a series of analyses on a particular set of data, there will be a proportion of statistically significant findings that are misleading 'false positives'.[4] A standard criticism of Bayesian confirmation theory is that it cannot represent the influence of biased methods on inference. We argue against this. One can use Bayesian confirmation theory to represent the influence of biased methods on confirmation.

P-hacking is related to a more traditional concern of philosophers, namely, the epistemic merit of the accommodation of existing evidence by a hypothesis which is tuned to accommodate that evidence versus the prediction of new evidence by a hypothesis which was developed prior to the observation of the evidence. Some have argued that prediction provides more confirmation than accommodation, while others have argued that prediction and accommodation are confirmationally equivalent.[5] This debate has more-or-less settled that prediction is epistemically superior to accommodation in

[2] In the US, a 1997 act reforming the Food and Drug Administration's use of clinical trials established an online web registry, sponsored by the National Institutes of Health, where medical researchers could register their pre-analysis plans (FDA, 1997). As of 2005, the most prominent journals in medicine require registration of clinical trials. While pre-analysis plans have been widely used in the medical sciences, they are relatively new in the social sciences. Further journal policies have emerged around pre-analysis plans in the social sciences. For example, several psychology journals now require pre-registration. Pre-analysis plans have also led to the rise of "results-blind review," in which journals commit to publishing a final paper if the analyses therein align with the analyses described in the plan. Several psychology journals have adopted this practice (Chambers (2013); Chambers, Feredoes, Muthukumaraswamy et al. (2014); Nosek, Lakens (2014)), and at least one journal in political science (*Comparative Political Studies*, see Findley, Jensen, Malesky et al. (2016)) and one in economics (*Journal of Development Economics*, see Foster, Karlan, Miguel (2018) for editorial statement) have followed suit.

[3] See Olken (2015); Coffman, Niederle (2015); Christensen, Miguel (2018) for summaries of some advantages and disadvantages of pre-analysis plans in the social sciences.

[4] Two influential articulations of this view are Leamer (1983) and Ioannidis (2005), in economics and medicine, respectively.

[5] For an overview of the debate, see Barnes (2008).

some—but not all—circumstances. We help ourselves to the fruits of this debate, and argue that some forms of p-hacking involve accommodation, which entails that, in some circumstances (but not all), evidence from p-hacking provides less confirmation than equivalent evidence that did not arise from p-hacking. Pre-analysis plans serve a function like prediction, thereby delivering epistemic benefits in particular circumstances.

This cluster of issues forms what we call in this paper the *problem of new evidence*. While the 'problem of old evidence' forced Bayesians to reckon with the implications of using existing data for confirmation of hypotheses, the problem of new evidence arises from a wholly different aspect of scientific discovery. In an era of 'big data,' vast computing clusters, techniques like meta-analysis, and powerful and easy-to-use statistical software, the generation of new evidence is cheap and plentiful (Leonelli, 2016). As such, theorists of scientific inference have to reckon with problems of new evidence—chief among them are the epistemic consequences of p-hacking and the ways in which particular methodological safeguards modulate those consequences. The contribution of this paper is thus to move beyond standard frequentist analyses of empirical scientific practice by providing a Bayesian articulation and analysis of the problem of new evidence.

We begin by describing the peril of p-hacking by appealing to the prediction-accommodation debate, Bayesianism, and model selection theory (§2). This contribution is novel, because the perils of p-hacking are almost exclusively discussed in a context of frequentist statistics. We describe the precise formal conditions under which p-hacking decreases confirmation, and conversely, the conditions under which p-hacking increases confirmation. We demonstrate that, contrary to the status quo, in some cases p-hacking can in fact be truth-conducive. The second contribution is an articulation of the epistemic benefits of pre-analysis plans, in which scientists note in advance what analyses they will perform as a means to mitigate the perils of p-hacking (§3). Even when such plans are in place, scientists often depart from them. In light of this fact, our third contribution is to describe the conditions under which such departures are dubious, and conversely, the conditions under which such departures are welcome (§4). §5 concludes.

§2 The perils of p-hacking

We motivate our analysis with a series of scenarios intended to highlight some puzzling and seemingly contradictory intuitions expressed in the ongoing discourse around p-hacking. Because some elements of p-hacking appear equivalent to the accommodation of existing data by tuning a hypothesis, we begin our analysis by drawing on the prediction-accommodation debate, before turning to our formal approach.

Consider the following two scenarios.

> SCENARIO 1
> A scientist wants to test the capacity of a drug to decrease the probability of heart attacks. She designs an experiment with all the proper methodological safeguards. Once the trial is over, she analyses the data and finds that the group that received the drug had the same frequency of heart attacks as the group that received a placebo.

She re-analyses the data by stratifying the population into males and females, and again finds no difference in frequency of heart attacks between drug group and placebo group, in neither males nor females. She then stratifies the population by age into two bins (young and old), and again finds no difference. But, when she combines the age and sex stratification, she finally observes that young men in the drug group had a lower risk of heart attack compared to the same demographic in the placebo group. Let E_1 be the evidence from the last analysis. Let H be the hypothesis 'this drug lowers the chance of heart attacks in young men.'

The precise description of "the evidence" E_1 will become a key focus of the discussion that follows. For the sake argument, for now, consider E_1 as constituted by the sample size, the mean treatment effects, and the distributional properties of the treatment effects.

SCENARIO 2
This scenario is just like Scenario 1 with two modifications: The scientist specifies H in advance, and the trial includes only young male subjects. E_2 is the analysis from this trial. E_2 has the same sample size, treatment effects, and distributional properties of treatment effects as E_1.

Scenario 1 involves p-hacking in virtue of the unconstrained subgroup analyses that the scientist performed. To render our analysis later in the paper tractable, we keep the volume of p-hacking and the range of p-hacking tactics in Scenario 1 very modest. As discussed in §1, p-hacking encompasses a range of practices. P-hacking can involve *experimental* tactics—by making multiple kinds of measurements on multiple properties of subjects, for example. P-hacking can also involve *analytic* tactics—by performing multiple statistical analyses on the same data, for example, or by binning the data in particular ways, such as occurs in Scenario 1. Further, p-hacking may include *publication* tactics—by only publishing subsets of one's evidence, for example. For now, we focus on the analytic and experimental tactics to isolate the issues at play and turn to reporting tactics in later sections.

An intuition that many hold about the epistemic peril of p-hacking is that the evidence in Scenario 1 provides less confirmation to the hypothesis than does the evidence in Scenario 2. To make this intuition more concrete, we evaluate the confirmatory power of the evidence in the two scenarios from the perspective of the scientific community. What are the justified credences of this epistemic community? For now, we assume that individuals in this community have full knowledge of the evidence-generating procedures—they know everything that the scientist did.[6]

One way of describing the intuition about the epistemic peril of p-hacking is with the following provisional statement:

[6] Again, this assumption allows us to isolate the methodological aspects of p-hacking, separate from the issues that arise purely from a lack of transparency.

(*p-hack*) If E_1 and E_2 represent equivalent data, E_1 and E_2 provide at least some confirmation to H, and E_1 is a result of p-hacking while E_2 is not, then $P(H \mid E_1) < P(H \mid E_2)$.

Though not articulated in precisely these terms, claims that rest on this basic intuition strike some as obviously true, and others as false. Our aim is to show that each view can be correct, sometimes, and to articulate the precise conditions under which *p-hack* is true. The main argument in our account below will involve a consideration that may have already occurred to readers, namely, that even if E_1 and E_2 represent equivalent data, a full description of what the 'evidence' is in the two scenarios entails that E_1 and E_2 are not equivalent. First, though, we attempt to understand the two scenarios above through the lens of the prediction-accommodation debate.

2.1. Prediction vs. accommodation

The precise problem with Scenario 1 is subtle. One approach to articulating the epistemic difference between Scenario 1 and Scenario 2 holds that E_1 was accommodated by H in Scenario 1 whereas E_2 was predicted by H in Scenario 2. To hold that Scenario 1 is epistemically inferior to Scenario 2 solely on the grounds that Scenario 1 involves accommodation of evidence whereas Scenario 2 involves prediction of that same evidence assumes that accommodation is generally inferior to prediction. However, accommodated evidence is not always epistemically inferior to predicted evidence, as the literature on the prediction-accommodation debate seems to have settled (Maher, 1988; Howson and Franklin, 1991; White, 2003; Barnes, 2008; Douglas and Magnus, 2013). This literature is vast, and addressing it adequately would take us astray, but to get a feel for the current thinking, consider the following scenarios.

SCENARIO 3
You toss a coin one thousand times. It lands heads on 503 tosses; that is your evidence (E_3). You then formulate a hypothesis H', which states that this coin is fair.

SCENARIO 4
You formulate a hypothesis H' which states that this coin is fair. You then toss the coin 1000 times. It lands heads on 503 tosses; that is your evidence (E_4).

We take it as intuitive that there is no epistemic distinction between these two scenarios.[7] That is, sometimes (perhaps often) there is an accommodation-prediction equality, as illustrated in Scenarios 3 and 4:

$$P(H' \mid E_3) = P(H' \mid E_4) \tag{1}$$

[7] The coin toss example is discussed in Maher (1988). A similar example is treated in Worrall (2014) and Frisch (2015).

Return for a moment to Scenario 1. Notice that E_1 could not possibly have falsified H, since H was formulated precisely to accommodate E_1. One might be tempted to hold that this is the problem with p-hacking: the evidence that results from p-hacking cannot possibly falsify a hypothesis that is formulated after the fact to accommodate that evidence. However, empirical scenarios in which (1) holds (that is, scenarios in which there is an accommodation-prediction equality) are such that the capacity to falsify the hypothesis is irrelevant. In Scenario 3, E_3 could not possibly have falsified H', but regardless, E_3 provides just as much confirmation to H' as does E_4.

If p-hacking is supposed to be a problem because the evidence that results from p-hacking is accommodated by a hypothesis rather than predicted by the hypothesis, then cases in which there is an accommodation-prediction equality suffice to show that, at the very least, p-hacking is not always a problem. However, in some cases prediction of evidence does provide more confirmation to a hypothesis than mere accommodation of the same evidence, and it is precisely this fact, one might say, that explains why some forms of p-hacking are at least sometimes epistemically pernicious. To see that prediction of evidence can provide more confirmation to a hypothesis than mere accommodation of the same evidence, consider the following scenarios.

SCENARIO 5
Mary wins the lottery (E_5). Then Joe says "the lottery is rigged in Mary's favour." Call Joe's hypothesis H''.

SCENARIO 6
Joe says "the lottery is rigged in Mary's favour" (H''). Then Mary wins the lottery (E_6).

We take it as intuitive that there is a significant epistemic difference between these two scenarios. That is, sometimes (perhaps often) there is an accommodation-prediction inequality, as illustrated by the contrast between Scenarios 5 and 6:

$$P(H'' \mid E_5) < P(H'' \mid E_6) \tag{2}$$

What distinguishes cases in which there is an accommodation-prediction equality from cases in which there is an accommodation-prediction inequality? Under what circumstances does prediction of evidence provide more confirmation than accommodation of the same evidence? In other words, when are we in scenarios represented by (1), and when are we in scenarios represented by (2)? A view that we find compelling is that in some cases, prediction of evidence suggests that the predictor has access to knowledge of a mechanism by which the evidence will be generated (Worrall, 2014; Frisch, 2015).[8] In Scenario 5, Joe has sour grapes, while in Scenario 6, Joe's astonishing successful

[8] We do not necessarily mean "mechanism" in the narrow and technical sense often deployed in philosophy of science, but rather, any plausible way in which the phenomenon in question could be brought about, which could involve an appeal to laws of nature, theories, causal regularities, or simple facts about the situation in question which explain the phenomenon. In Scenario 6, for example, perhaps Joe saw how the lottery operator rigged the chance device in favour of Mary.

prediction suggests that he knows something about the machinations of the lottery.[9] In contrast, there is no such difference between Scenario 3 and Scenario 4: everyone is familiar with fair coins and most of us have a passable understanding of the physics of coin fairness. There is no epistemic difference between the accommodation in Scenario 3 and the prediction in Scenario 4 because in both scenarios knowledge of the underlying evidence-generating mechanism is rich.

Now, let us return to p-hacking. In Scenario 2, the fact that the scientist specifies H in advance, and E_2 ends up confirming H, suggests that the scientist has a decently--founded speculation about the mechanism by which the drug operates. This mechanism could explain why the drug is effective only for a particular demographic group. In Scenario 1, by contrast, the scientist's shotgun approach to analysis suggests that she has no knowledge of such a mechanism.

In short, *p-hack* might be a result of the distinction between prediction and accommodation, which itself can be understood as a result of knowledge of underlying mechanisms which produced the evidence, in some cases (like Scenario 6 compared with Scenario 5), but not in other cases (like Scenario 4 compared with Scenario 3). Cases of p-hacking provide less confirmation than their non-p-hacking equivalents, in some cases, because p-hacking involves accommodation while their non-p-hacking equivalents involve prediction, and in some cases, prediction looks like Scenario 6 and accommodation looks like Scenario 5. In scenarios like 5 and 6, *p-hack* holds, while in scenarios like 3 and 4, *p-hack* does not hold.

2.2. Bayesian confirmation theory

The appeal to knowledge of mechanisms to explain the epistemic difference between prediction and accommodation is intuitive, though perhaps a tad opaque. While the discussion of prediction and accommodation helps to plumb our intuitions about p-hacking, it also highlights the importance of further precision. In particular, the preceding discussion remains relatively vague about what comprises "evidence" in different scenarios, and under what circumstances such evidence provides confirmation to hypotheses of interest. Further, as a general rule it is compelling to explicate local principles of scientific inference, such as *p-hack*, by appealing to a more fundamental and more general theory of scientific inference. In what follows we explicate the perils of p-hacking using Bayesian confirmation theory.

This might strike knowledgeable readers as an impossible task. One way to understand the difference between Scenario 1 and Scenario 2 is that in Scenario 2 the scientist has a pre-specified 'stopping rule' (an analysis plan) while in Scenario 1 there is no such stopping rule. We do not have the space to explain this issue to the uninitiated,

[9] We stipulate the evidence in Scenarios 5 and 6 as "Mary wins the lottery." Though we will get into a more thorough discussion of the description-sensitivity of evidence below, note that there is another way of describing E_5 and E_6 that yields accommodation-prediction inequality. If E_5 is "Mary wins the lottery *before* Joe forms his beliefs about the fairness of the lottery" and E_6 is "Mary wins the lottery *after* Joe forms his belief about the fairness of the lottery," then Joe's mechanistic knowledge is foregrounded in E_6 itself, and (2) intuitively follows.

but note simply that frequentist statisticians and philosophers claim that stopping rules are epistemically relevant, while many Bayesians deny this. Indeed, critics of Bayesianism claim that features of experimental design (such as presence or absence of p-hacking) are epistemically irrelevant once data has been collected (Mayo, 1996). So, since Bayesians claim that stopping rules are epistemically irrelevant, and *p-hack* assumes the epistemic relevance of stopping rules, Bayesians seem committed to outright denying *p-hack*. On the face of it, Bayesians cannot articulate the perils of p-hacking, because they do not believe there is a peril. This is all wrong. Bayesians can and should articulate the perils of p-hacking. That is what we do now.

A Bayesian can explain *p-hack* in a limited number of ways. By Bayes' Theorem, the inequality in *p-hack* is equivalent to:

$$\frac{P(E_1 \mid H)\ P(H)}{P(E_1)} < \frac{P(E_2 \mid H)\ P(H)}{P(E_2)} \qquad (p\text{-}hack')$$

This inequality cannot result from a difference in the prior probability of the hypotheses $P(H)$, because, obviously, the hypothesis is the same. Indeed, $P(H)$ simply drops out of *p-hack'* since it appears on both sides of the inequality in *p-hack'*. Since the inequality in *p-hack'* is not a result of a difference in the priors, it must be a result of a difference in the likelihoods or the expectancies of the evidence.

Because E_1 and E_2 appear to be identical, one might think that their expectancies, $P(E_1)$ and $P(E_2)$ must be the same. Moreover, the familiar 'problem of old evidence' suggests that since both E_1 and E_2 are stipulated in the scenarios as given, their probabilities are both 1. Glymour (1980) raised this as a problem for Bayesianism because if the expectancy of evidence is 1, then that evidence cannot increase the probability of a hypothesis.[10] Both problems suggest that *p-hack'* cannot be a result of a difference between the expectancies of the evidence, $P(E_1)$ and $P(E_2)$. However, a standard solution to the second problem happily serves as a solution to the first problem, and moreover, provides us with the insight we need to explain *p-hack*.

The solution to the problem of old evidence is to ask, counterfactually, what the expectancy of the evidence would have been, prior to being given the evidence. In Scenario 2, for example, once the experiment is over and the scientist has E_2, one ought not determine $P(E_2)$ by asking what the actual probability of E_2 is, since, at that point, the actual probability of E_2 is 1, and thus the problem of old evidence arises. Rather, one ought to determine $P(E_2)$ by asking what the counterfactual probability of E_2 would have been, given all background knowledge excluding E_2.[11] For the purpose of understanding *p-hack* we do not need to know the precise values of $P(E_1)$ and $P(E_2)$. Instead, we merely need to show:

[10] If E is known, then $P(E) = P(E \mid H) = 1$. Then, by Bayes' Theorem, $P(H \mid E) = P(H)$, and thus the evidence does not confirm the hypothesis.

[11] Glymour himself suggested this solution, but dismissed it on the grounds that such counterfactuals are beyond the insight of scientists; Howson (1991) and other Bayesian are unmoved by this complaint; for a recent application of this approach, see Frisch (2015).

$$P(E_1) > P(E_2) \qquad (*)$$

If (*) holds, then so does *p-hack'*, and we have an explanation of the perils of p-hacking.

Up to this point we have been vague about the sense in which E_1 and E_2 are equivalent. Here, we put pressure on this vagueness and show that it is precisely the description of the evidence that generates the inequality in *p-hack'*. Described a certain way, the expectancies $P(E_1)$ and $P(E_2)$ are equal. Recall that both E_1 and E_2 are 'young men in the drug group had a lower risk of heart attack compared to the same demographic in the placebo group', with the same sample size, and same effect size (mean and distribution). At this level of detail, E_1 and E_2 are identical and thus their expectancies must be identical. Changing the description of the evidence in Scenario 1, however, increases the expectancy of the evidence.

To see this, ask: what would be the probability that we would observe evidence like *that*? Fairly high, or at least, higher than the probability that we would observe that same evidence in Scenario 2. The reason here is both simple and subtle. The simple reason appeals to the 'law of truly large numbers', which holds that "with a large enough sample, any outrageous thing is likely to happen" (Diaconis and Mosteller, 1989, as cited in Sober, 2008). Each subgroup analysis in Scenario 1 can be thought of as another sampling of the data: the more such analyses are performed, the larger the sample. Consider Tamara, who was struck by lightning twice. What are the chances of that?! The 'law of truly large numbers' entails: someone is bound to be hit by lightning, and with enough people around and after enough time passes someone is bound to be hit by lightning twice; that person just happened to be Tamara. Similarly, some subgroup analysis is bound to report a positive effect, if enough analyses are performed. The larger the sample, the more likely it is that an outrageous thing will happen. To understand what 'outrageous' means, however, we need the subtle reason.

The subtle reason notes that the expectancy of the evidence is sensitive to the description of the evidence. Suppose that E_1' is described as the conjunction of E_1 with knowledge that the other subgroup analyses were performed. This entails that $P(E_1')$ must be less than or equal to $P(E_1)$ and thus less than or equal to $P(E_2)$, and thus (*) would not be satisfied and we could not make sense of p-hack'.[12] Perhaps E_1 should be redescribed as E_1'': "multiple subgroup analyses were performed and in one of them there was a positive effect." Return to the lightning analogy: the probability of *Tamara* being struck by lightning twice is extremely low; the probability of *someone* being struck by lightning twice is extremely high. Under this description we do not specify which subgroup analysis had the positive effect, and thus we can apply the 'law of truly large numbers', and conclude that $P(E_1'') > P(E_1)$ and thus $P(E_1'') > P(E_2)$. This would therefore allow us to explain *p-hack*.

One might object that describing the evidence from Scenario 1 as E_1'' involves discarding information, and therefore violates the principle of total evidence. Moreover, as Sober (2008) argues, having the freedom to weaken the description of one's evidence this

[12] This entailment follows directly from E_1' being the conjunction of E_1' with other knowledge. The conjunction (E_1') cannot have a higher probability than one of its conjuncts (E_1).

way simply to align the evidence with one's pre-theoretic views runs the risk of using an evidence-weakening strategy simply to get the evidence to fit one's view. However, the principle of total evidence does not prohibit all cases of logically weakened evidence, as the example of Tamara being struck twice by lightning suggests. Sometimes logically weakening one's evidence is reasonable, given what one knows about the context in which the evidence was generated. It is intuitive that a description of one's evidence can include too much information. Suppose that in Scenarios 3 and 4, the description of the evidence included the colour of your socks, your zodiac sign, and the rain forecast for Tahiti. E_3' and E_4' would thus be "the coin landed heads 503 times, and the coin-tosser is a Scorpio wearing blue socks, and the probability of rain in Tahiti is 20%." What makes the additional content in the description of the evidence excessive is the obvious fact that the additional content is irrelevant to the truth of the hypothesis. We make that judgement based on our background knowledge of, say, the absence of causal influence between the sock colour of a coin-tosser and the results of their coin tosses. These considerations apply to a description of any evidence: do we have reasons to think that the particular features articulated in the evidence description are relevant (perhaps causally) to the hypothesis under investigation? If yes, then those features should be included in the evidence description, and if no, then not.

We have argued that some cases of p-hacking entail an increased expectancy. Crucially, however, some cases of p-hacking can also contribute to a high likelihood, $P(E|H)$ because a scientist searches for evidence which fits her hypothesis or the hypothesis is tuned precisely to accommodate the evidence. This entails a high likelihood.

To sum: p-hacking contributes to a high likelihood, $P(E|H)$, and also contributes to a high expectancy of the evidence, $P(E)$. Since confirmation of a hypothesis is proportional to the ratio of the likelihood to the expectancy, and p-hacking increases both, there is nothing general to say about the impact of p-hacking on confirmation: some cases of p-hacking will increase the likelihood more than the expectancy, and other cases of p-hacking will increase the expectancy more than the likelihood. This influence of p-hacking on confirmation holds both when the hypothesis in question is true and when it is false. Thus, if p-hacking has a positive impact on confirmation but the hypothesis is false, p-hacking leads us astray, and similarly, if p-hacking has a negative impact on confirmation but the hypothesis is true, p-hacking leads us astray; in both cases p-hacking is epistemically pernicious. However, if p-hacking has a positive impact on confirmation and the hypothesis is true, p-hacking pushes our credence toward true belief, and similarly, if p-hacking has a negative impact on confirmation and the hypothesis is false, p-hacking pushes our credence toward true belief; in both cases p-hacking has guided us towards truth. The relative frequency of the latter cases may be small, but their existence suffices to show that some practices that constitute p-hacking are not universally misleading (it is worth repeating that we are here focused solely on the methodological aspects of p-hacking and not on the aspects of p-hacking that involve selective reporting of evidence).

2.3. Model selection theory

Some may have reservations about the Bayesian approach to understanding p-hacking, or to Bayesianism generally. The perils of p-hacking can also be articulated using the resources of 'model selection theory', most prominently applied to philosophical questions by Sober (see, for example, Sober (2008, 2015)). We follow Sober's approach, which draws on the Akaike framework, which holds that simpler models are more predictively accurate than complex models, and models with better fit-with-data are more predictively accurate than models with worse fit-with-data.

Briefly, model selection theory employs quantitative criteria to reward models with high likelihood and punish models with low parsimony, where parsimony is determined by the number of adjustable parameters in the model. One such approach is the Akaike Information Criterion, which, in the version offered by Sober (2015), is: $\log\{P[Data \mid L(M)]\} - k$. In this formulation, $\log\{P[Data \mid L(M)]\}$ represents the likelihood, or fit with data, of a fitted model M, and k represents the number of free parameters of the model. Fewer free parameters entails greater parsimony. The focus on predictive accuracy in model selection theory supplants the focus on credence in Bayesian confirmation theory. Arguably, predictive accuracy is what one is aiming at in cases like Scenario 1 and 2: we want to know if the results of the experiment give us reason to take the drug.

In Scenario 1, the scientist began by testing the effect of the drug on the population as a whole, and then proceeded to test the effect of the drug on the four subpopulations based on gender (male, female) and age (young, old). The overall population had an effect size p. The scientist's initial analysis of the data, prior to the subgroup analyses, was a test of the following model, M:

> M: There is a number x_0 such that $p = x_0$ and $x_0 > 0$.

In virtue of the scientist's single analysis of the full population, the model tested by the scientist has a single free parameter, x_0. In virtue of the fact that it is an experimental drug under investigation, and drugs are supposed to be helpful, the model holds that the effect size is positive, and thus $x_0 > 0$. M is a simple model, and so it scores well with respect to the parsimony aspect of model selection theory. M does not score well with respect to the fit-with-data aspect of model selection theory, because, as described in the scenario, p in fact is zero, and thus $x_0 = 0$, which fits poorly with M.

Each subgroup analysis tests whether the effect size of the intervention was different in that subgroup relative to the overall population effect size. Call the effect sizes in the four subgroups a, b, c, and d. Thus, the scientist's subgroup analyses amounted to a test of the following model, N:

> N: There is a number x_1 such that $x_1 = a$ and $x_1 \neq p$ and there is a number x_2 such that $x_2 = b$ and $x_2 \neq p$, and there is a number x_3 such that $x_3 = c$ and $x_3 \neq p$, and there is a number x_4 such that $x_4 = d$ and $x_4 \neq p$.

The model selection comparison of *M* and *N* is revealing. *N* has four adjustable parameters (x_1 – x_4), and *M* has only 1 (x_0). Of course, *N* has a higher fit-with-data score than *M,* because, as described in the scenario, one of the subgroup analyses had a positive effect size, and so one of the conjuncts in the model has good fit-with-data.

Because the scientist first tested *M,* and subsequently tested *N,* it is reasonable to describe the model that the scientist tested as:

$$M \,\&\, N$$

This model has five free parameters (x_0 – x_4), and so is inferior to *M* on parsimony, but is superior to *M* on fit-with-data.

What does this tell us about p-hacking? Just like the Bayesian analysis, the model selection argument shows that p-hacking is sometimes, but not always, epistemically pernicious. The influence of an instance of p-hacking on the goodness of a model (that is, the predictive accuracy of a model) depends on the relative influence of p-hacking on the parsimony score of the model versus the fit-with-data score of the model. P-hacking decreases the parsimony of models of empirical scenarios but increases the fit-with-data of models of empirical scenarios. If p-hacking decreases parsimony more than it increases fit with data, then p-hacking mitigates predictive accuracy, but if p-hacking increases fit with data more than it decreases parsimony, then p-hacking enhances predictive accuracy.

In this section, we have demonstrated the epistemic peril of p-hacking by drawing on the resources of predictivism, and on the formal tools of Bayesianism and model selection theory. The upshot is that p-hacking is sometimes, but not always, epistemically pernicious. We have described the precise formal conditions under which this is so.

§3 The promise of pre-analysis plans

One response to the threat of p-hacking in medical sciences and empirical social sciences is the requirement that scientists register their experimental and analytic plans in a public venue (such as a database or journal publication). In a pre-analysis plan, researchers specify in advance the measurements they plan to gather and the statistical analyses they plan to perform, committing themselves to specific hypotheses to be tested before gathering and analysing data. Pre-analysis plans are used widely in the medical sciences but have only begun to gain traction in empirical social science research in the last decade.[13] In this section we describe the epistemic function of pre-analysis plans, building on the formal analysis in the previous section. We focus on how pre-analysis plans address the *analytic* and *experimental* tactics described as p-hacking, and leave a discussion about how pre-analysis plans relate to *reporting* tactics to §4.

For a given experimental or observational method, a pre-analysis plan specifies the primary features to be measured (statisticians call this the 'primary outcome var-

[13] For discussions of the adoption of pre-analysis plans in the social sciences, see e.g. Coffman, Niederle (2015); Casey, Glennerster, Miguel (2012); Miguel, Camerer, Casey et al. (2014); Simmons, Nelson, Simonsohn (2011); Humphreys, Sanchez, Windt (2013); Brodeur, Lé, Sangnier et al. (2016).

iable'), other features to be measured ('secondary outcome variables'), rules for data inclusion and exclusion (to deal with flawed measurements and missing data), the subgroup analyses to be performed, and which statistical models and tests will be used. In a discussion of hypothetical pre-analysis plans for a randomised trial performed to test the effects of teacher monitoring programs on student performance, Olken (2015) notes the specificity required for pre-analysis plans:

> What test and test subjects are included? Will the outcome variable be the test score in levels or logs? Will it be in standard deviations, the percentile of the test score, a binary variable for passing the test, a binary variable for being above the 25th percentile, or the 50th percentile, and so on? Will the score be in levels or an improvement from a baseline? If there are multiple subjects, like math and reading, how will the scores be aggregated into a single outcome variable? Are there any rules for trimming or excluding outliers?

Such specificity is meant to minimise p-hacking. Without such specificity, a researcher could, say, gather data on many primary features and perform as many analyses as required to find a positive finding.

Thus, one function of pre-analysis plans is to minimise the analytic and experimental tactics that amount to p-hacking. If p-hacking is unreliable, and if an empirical method (construed broadly to include the planning and execution of the method, the analysis of the data, and the publication of the results) includes a design feature which minimises the extent of p-hacking, then that design feature enhances the reliability of that method.

On the other hand, the very same aspects of pre-analysis plans that constrain p-hacking also constrain the ability of researchers to explore data in the hope of discovering true findings. There might be an important fact of nature lurking in a data set, but if scientists are constrained from searching through the data then that fact of nature will not be then discovered. Pre-analysis plans might cause researchers to neglect unexpected findings. Pre-analysis plans place scientists far on one extreme end of the trade-off between avoiding false positive findings and avoiding false negative findings. This position regarding the trade off between false positive findings and false negative findings is not generally justified. It would be strange, for example, to prohibit the 'discovery' in Scenario 3. This is not merely an erudite issue, since many scientific research programmes today, such as big data methods in genome-wide association studies, require the unconstrained, post-hoc analysis of data to find correlations between genetic features and phenotypic features.[14]

There are straightforward arguments both for and against the use of pre-analysis plans. A well-founded articulation of the epistemic function of pre-analysis plans could explain the particular merits of the arguments on both sides. That is our ambition here. In §2 we articulated the problem with p-hacking by employing Bayesian confirmation

[14] See, for example, Pearson, Manolio (2008), who argue that "the GWA [genome-wide association] approach can be problematic because the massive number of statistical tests performed presents an unprecedented potential for false-positive results."

theory and model selection theory. Here we describe the epistemic function of pre-analysis plans by drawing on these two frameworks.

In short, the function of pre-analysis plans is simple: pre-analysis plans constrain the experimental and analytic freedom of scientists.[15] Those in favour of pre-analysis plans take this constraint to be epistemically meritorious precisely because such constraint mitigates the peril of p-hacking. We will call this the Pro argument. Those opposed to pre-analysis plans take such constraint to be epistemically nefarious precisely because such constraint mitigates the extent to which scientists can learn from data. We will call this the Con argument. Those opposed to pre-analysis plans add that pre-analysis plans do not have any epistemic merit; this view holds that the Pro argument is based on a misguided inductive (frequentist) framework. Let's call this the Anti-Pro argument. Since §2 was devoted to articulating the perils of p-hacking from inductive frameworks other than frequentism, we can dismiss the Anti-Pro argument: any serious philosophical theory of scientific methodology must recognise the peril of p-hacking, and any methodological device that minimises p-hacking must have at least some epistemic merit, at least in particular circumstances.

We are left with the Pro argument and the Con argument. Both seem compelling. They are, obviously, at odds.

However, we saw that the two arguments share a premise. They both hold that the epistemological function of pre-analysis plans is due to the constraint a pre-analysis plan places on scientists. It is worth noting that such constraint is not guaranteed by pre-analysis plans, and that such constraint can be achieved in other ways. Imagine a pre-analysis plan that allows or stipulates a great number of parameters to be measured in an experiment and a great number of analyses to be performed on such measurements, and scientists follow the plan, diligently trolling through data as the plan stipulates. This pre-analysis plan offers little constraint. Now imagine a scientist who sets out to perform an experiment with no pre-analysis plan, but her resources are extremely limited: she only has time and money to measure a single parameter and analyse the data once. In the absence of a pre-analysis plan, this scientist is constrained to the same extent as, say, the scientist in Scenario 2. Nevertheless, in practice pre-analysis plans serve to constrain the methodological and analytic freedom of scientists. Standard pre-analysis plans are not as loose as this imaginary one, and they are often employed in empirical scenarios in which multiple analyses are not otherwise constrained (in many scenarios multiple analyses are cheap and easy to perform).

The Pro argument is directly warranted by noting the perils of p-hacking: the same considerations that we raised in §2 that show that p-hacking is epistemically pernicious entail that any methodological tactic that mitigates p-hacking can provide some epistemic benefit. Thus, from both the perspective of Bayesianism and model selection theory, the Pro argument gets some vindication. Pre-analysis plans are epistemically meritorious precisely because they mitigate p-hacking.

[15] Again, note that pre-analysis plans may serve other purposes beyond their role in constraining experimental and analytic freedom. For example, they may mitigate selective reporting tactics that are also described as p-hacking. We will discuss these issues in §4.

More specifically, since the Bayesian articulation of the peril of p-hacking holds that p-hacking increases $P(E)$ and since pre-analysis plans mitigate p-hacking, it follows that pre-analysis plans can serve to decrease $P(E)$. This, in turn, adds confirmation to one's hypothesis (via Bayes' Theorem).

Similarly, since our model selection theory articulation of the peril of p-hacking holds that p-hacking increases the number of free parameters of a model tested in an empirical scenario, and since pre-analysis plans mitigate p-hacking, it follows that pre-analysis plans can serve to decrease the number of free parameters of this model. This, in turn, increases the predictive accuracy of the model (via the Akaike Information Criterion).

Vindication for the Con argument is also straightforward. Consider first the perspective of Bayesianism. Recall that the Con argument claims that pre-analysis plans unduly mitigate the extent to which scientists can learn from data. For a Bayesian, learning from data is represented by the likelihood, $P(E \mid H)$. Suppose, as in Scenario 1, we allow a free exploration of data from an experiment, and a positive signal (E) is found in one of the analyses, and then a hypothesis (H) is formulated to 'accommodate' that evidence. The likelihood is, therefore, high (arguably, the likelihood is 1, or very close to 1, because the hypothesis was formulated precisely to accommodate the evidence). If this free exploration of data were constrained by a pre-analysis plan, it is possible that this particular signal would not have been noticed or this particular hypothesis would not have been pre-specified; the evidence that a scientist would get in this constrained scenario would have a lower likelihood. Since high likelihoods are a desideratum in science, this is an argument for permitting the free exploration of data, and thus an argument against the use of pre-analysis plans. This is a Bayesian vindication of the Con argument.

Consider now the model selection vindication of the Con argument. It is virtually identical to the Bayesian vindication of the Con argument. We saw in the prior paragraph that allowing free exploration of data tends to generate higher likelihoods than would be case were analysis of data constrained by pre-analysis plans. Since 'fit-with-data' is rewarded in model selection theory, and at least in the Akaike framework this is formulated with the likelihood, it follows that pre-analysis plans will contribute to lower 'fit-with-data' (likelihoods), and thus contribute to poorer predictive accuracy of a chosen model.

The Bayesian vindication of the Pro argument holds that pre-analysis plans decrease $P(E)$. The Bayesian vindication of the Con argument holds that pre-analysis plans decrease $P(E \mid H)$. With this in place it is easy to see how to resolve the two arguments. A hypothesis gains confirmation if $P(E \mid H) > P(E)$. Thus, supposing H is true, a pre-analysis plan is epistemically meritorious if and only if its influence on decreasing the expectancy of the evidence, $P(E)$, outweighs its influence on decreasing the likelihood, $P(E \mid H)$. Conversely, when H is false, a pre-analysis plan is epistemically meritorious if and only if its influence on decreasing the likelihood outweighs its influence on decreasing the expectancy of the evidence.

The model selection vindication of the Pro argument holds that pre-analysis plans increase parsimony of models of empirical scenarios, and thus increase predictive accuracy of those models. The model selection vindication of the Con argument

holds that pre-analysis plans decrease fit-with-data (likelihoods) of models of empirical scenarios, and thus decrease predictive accuracy of those models. With this in place it is now easy to see how to resolve the arguments. A pre-analysis plan is epistemically meritorious if and only if its influence on increasing the parsimony of a model of an empirical scenario outweighs its influence on decreasing the likelihood of that model. Thus, from both the Bayesian perspective and the model selection perspective, one cannot say that pre-analysis plans are in general epistemically meritorious or nefarious. We have, however, succeeded in showing the precise formal conditions under which pre-analysis plans are epistemically meritorious and, conversely, the precise conditions under which pre-analysis plans are epistemically nefarious.

The illumination of this core feature of contemporary experimental design is valuable in itself. More than this, though, we can now begin to understand a ubiquitous and, at first glance, worrying aspect of scientific practice. Even when pre-analysis plans are in place, scientists often depart from them. Second-order empirical studies suggest that departures from pre-analysis plans are widespread. We turn now to this apparent problem.

§4 Departures from plans

Thus far we have discussed the epistemic threat of p-hacking and the role of pre-analysis plans in mitigating that threat when it is present. Often, however, pre-analysis plans are in place and yet scientists nevertheless change their analysis strategy, thereby departing from the proposed plan (Dwan et al., 2008). How should we assess the epistemic merit of a scientific study under such circumstances? In this section, we begin to examine the epistemic role of pre-analysis plans when the plans are not followed strictly. Thus far we have discussed the *analytic* and *experimental* tactics that may be described as p-hacking, and how pre-analysis plans may address them. But, another set of tactics thus far overlooked are *reporting* tactics often described as p-hacking.[16] Departures from plans highlight the interplay between what we have called analytic tactics and reporting tactics. If a plan is "flexible" in the sense that some departures are allowed, it weakens the analytic constraints on researchers. But, the existence of the plan places higher demands on researcher transparency, and thus constrains the degree of selective reporting.

Our motivation for considering departures from pre-analysis plans is twofold. First, no pre-analysis plan can be completely exhaustive in its description of the scientific investigation to be undertaken. So, it is important to better understand the ways in which the epistemic merits of a pre-analysis plan depend on its specificity and a researcher's faithfulness to it. Second, we are motivated to study departures from plans because of how pre-analysis plans are often used in practice. Researchers depart from their registered plans for a variety of reasons stemming from both practical and conceptual vagaries of scientific practice.

[16] We have ignored selective reporting of results until now because such practices are more obviously problematic than the analytic issues that took centre stage above—if a researcher intentionally misrepresents the set of analyses performed, this untruthfulness is likely to lead the epistemic community further from the truth.

To ground our discussion in the realities of scientific research, we look to a landmark use of a pre-analysis plan in the social sciences. Development economists Katherine Casey, Edward Miguel and Rachel Glennerster set out to evaluate a particular attempt to "reshape institutions" to make them more democratic and egalitarian (Casey et al., 2012). They analyse the results from a randomised trial of a governance program in Sierra Leone, and discuss how the use of a pre-analysis plan can "help avoid some common pitfalls in empirical research." What statistical analyses would suggest that the intervention did or did not succeed?

Before the intervention began in 2005, the authors made a list of three hypotheses they were interested in investigating. When the intervention ended in 2009, but before they began their analyses, they augmented the list with eight further hypotheses. When they began their econometric analyses, they had a list of eleven hypotheses, and while writing the article they added a twelfth. They refer to the twelve hypotheses under consideration H_1–H_{12}.[17] In their analyses, they present evidence for each of the hypotheses, taking extra care to note which ones were specified in advance.[18]

Casey et al.'s experience leads them to a position that pre-analysis plans are most useful when they are *flexible*—that is, they can be epistemically useful even if they are not strictly followed. In their words:

> We advocate a compromise position that allows some researcher flexibility accompanied by the "price tag" of full transparency—including a paper trail of exactly what in the analysis was prespecified and when, and public release of data so that other scholars can replicate the analysis—with the hope that this approach will foster the greatest research progress.

Is their view compatible with the preceding analyses? We argue that it is. We vindicate a version of the view informally articulated by Casey et al. (2012). As such, and in light of the preceding discussions, we take flexible pre-analysis plans to be a useful—if incomplete—solution to the "problem of new evidence."

4.1. Analytic constraints

What happens when a researcher who has registered a plan chooses not to follow the plan? A tempting answer might be: the plan is not valuable because it was not strictly followed. Our argument in §3 made the assumption that when a researcher uses a plan, she follows it. This tempting response holds that when a researcher chooses not to follow the pre-analysis plan, it is as if she did not make a plan at all.

But we take this answer, tempting as it may be, to be overly simplistic. A plan can be epistemically relevant even if it is not strictly followed. We show that—compared to a setting in which there is no plan at all—a plan can offer *some* analytic constraints. Though these analytic constraints are not binding, they still may be epistemically relevant.

[17] For a complete list of the hypotheses they investigated, see Appendix.

[18] They correct for multiple hypothesis testing with Bonferroni corrections. As Bonferroni corrections largely aim to mitigate the frequentist's issue with p-hacking, we do not discuss them here.

Recall the conclusion of our Bayesian analysis of pre-analysis plans. They decrease $P(E)$ which is an argument in their favour (supposing H is true), but they increase $P(E \mid H)$, which is an argument against their use. To suggest that a flexible plan has no merit is to suggest that neither the expectancy nor the likelihood could change if a plan is not strictly followed. But we have offered reasons to think that this is not the case.

Clearly, having some analytic constraints may be better than having none at all. But at the same time, to consider whether pre-analysis plans are valuable, we need to understand the tradeoff between constraining the freedom of researchers and decreasing false positives. In some cases, *flexible* pre-analysis plans could function as a happy medium between the two extreme cases— they may still decrease $P(E)$ while increasing $P(E \mid H)$ less than a rigid plan would.

Consider, for example, settings in which the multidimensionality of the outcomes of interest make it difficult for a researcher (and therefore the epistemic community) to decide how to run their analyses. In the Casey et al. (2012) case, defining the outcome of interest—whether the intervention "strengthened institutions"—is hard to do ex-ante. It is useful to see the definition of the outcomes as a process, which involves letting the data speak without cherry picking outcomes. As they write:

> The multidimensionality of institutions — governing political, economic, and social behaviours— implies a large number of outcomes … some of which will have statistically significant treatment effects by pure chance. Moreover, because institutions are amorphous and contextually determined, there is no commonly agreed-on set of standard measures defining the core of each domain, allowing the researcher to either deliberately or unintentionally "cherry pick" a set of treatment effects whose selectivity is difficult to detect from the outside.

When there is a plan in place—even if it is not strictly followed—the selectivity may be more guided by scientific reasoning than by chance. The plan offers some discipline for the researchers in choosing which analyses to pursue, decreasing $P(E)$ compared to a case with no plan. But the plan, if it is flexible, can lead to an increase in $P(E \mid H)$ relative to a comparison case in which the plan is rigid. Thus, flexible plans are an appealing policy option in this middle ground. Flexible plans may retain the benefits of a rigid plan, while offering researchers more analytic constraint than in parallel cases in which no plan is pre-specified.

4.2. Transparency

Thus far, all of our analyses have focused on *analytic* and *experimental* tactics that amount to p-hacking, and how pre-analysis plans may mitigate the adverse effects of p-hacking on scientific discovery. Now, we discuss how *flexible* pre-analysis plans address *reporting* tactics that may amount to p-hacking.

A departure from a plan implicitly or explicitly involves the formulation of a new hypothesis. Casey et al. (2012) make their departure very clear by noting which hypotheses were generated before the intervention, before any analyses, and while analyses were

ongoing. Call the initial hypothesis (or set of hypotheses) H and the modified hypothesis (or set of hypotheses) H'. It is possible that having *some* insight into the methods of the researcher lends greater confirmation than a complete black-box as to how the hypothesis H' came about. When there is no plan at all, the epistemic community has no information about why the researchers chose to run some analyses and not others.

We represent the existence of a plan corresponding to H by C. The departure from the plan constitutes *second-order evidence*—evidence about how the evidence was generated. Suppose a researcher departs from a pre-analysis plan, and receives some confirming evidence E'. We are now interested in the probability that H' is true given E', but also given the other information available. The researcher does not have a pre-analysis plan that corresponds to the hypothesis H'. But the researcher does have a plan for a different hypothesis, H, so we have C. The conditional probability we are interested in, then, is:

$$P(H' \mid E', C) \tag{3}$$

When there is a plan, the epistemic community may have the ability to evaluate the juxtaposition of H and H' and determine that the extent of p-hacking is bounded. These bounds on the extent of p-hacking then constitute background knowledge that can influence the degree of confirmation of the evidence on the hypothesis. In the no plan case, there is no upper limit to the possible extent of p-hacking. But, when there is a plan in place, the juxtaposition of the hypothesis H and H' may provide some bounds on the extent of p-hacking that may have occurred.

4.3. Further issues

We have not considered here the opportunities for strategic manipulation of flexible policies. A major challenge to our argument in support of "flexible" pre-analysis plans is that such a norm or policy would simply alter the strategic incentives of researchers wishing to p-hack. Rather than having incentives to p-hack, researchers would have incentives to write pre-analysis plans and create the appearance of transparency. Nefarious researchers could make efforts to give the impression of full transparency while in actual fact they offer only partial insight into the ways in which they departed from their stated plans. This appearance of transparency could serve to mask p-hacking, and perhaps make it even less detectable than in the absence of a plan. Such dynamic incentive issues are beyond the scope of this paper, but certainly deserving of further study. In this vein, several recent models take a mechanism design approach to optimal research transparency policies (Frankel and Kasy, 2020; Libgober, forthcoming), while others look at the virtues of various scientific norms around transparency from a social epistemology perspective (Bright, 2017).

In §3, we offered a modest defence of pre-registration. In this section, we showed that there are cases in which a pre-analysis plan may be epistemically relevant even if it is not strictly followed. This analysis points to a potential role for pre-analysis plans that breaks from their stated purpose. Rather than serving as a constraint—and only a constraint—for researchers, they might serve as a tool for greater transparency in the

epistemic community writ large. Pre-registration can lend greater insight to researchers' methods, even if the researchers' investigations do not perfectly align with the pre-registered plan. This increase in transparency due to pre-registration can be epistemically influential in a number of ways. Recognising this potential role for pre-registration—as a tool for providing transparency rather than constraining researchers—could lead to improvements in scientific practice. More generally, it offers a way of thinking about the 'problem of new evidence' associated with p-hacking.

§5 Conclusion

We have argued that experimental and analytic tactics that amount to p-hacking are sometimes but not always epistemically pernicious. We made this argument by drawing on the fruits of the prediction-accommodation debate, and by appealing to formal tools from Bayesian confirmation theory and model selection theory. In accordance with our position on p-hacking, we articulated a modest defence of pre-registration. Finally, we suggested—counterintuitively—that pre-registration may be epistemically relevant regardless of whether the pre-registered plan is strictly followed. Thus, *flexible* pre-analysis plans may be an appealing tool for epistemic communities to confront what we have called the *problem of new evidence*.

Our argument has philosophical and practical implications. P-hacking is a lightning rod for a cluster of related issues in philosophy of science: prediction vs. accommodation, stopping rules, values in science, and more. We have shown that p-hacking can be understood as a real-world instantiation of the prediction-accommodation debate. Furthermore, we demonstrated the value of Bayesian confirmation theory and model selection theory in understanding p-hacking and pre-registration.

On the practical side, p-hacking has been maligned widely as a cause of the replication crisis, and pre-registration has emerged as an important methodological safeguard against it. Our analysis suggests that this view of p-hacking—and the corresponding view of pre-registration— is too coarse. As p-hacking is not always epistemically pernicious, pre-registration best serves as an opportunity for increased transparency around researcher methods, rather than as a strict methodological bind preventing exploratory analysis. While pre-registration should continue to be encouraged, the purpose of pre-analysis plans must be clarified and revised in order to align with the subtleties of p-hacking presented here.[19]

[19] For discussion of this paper we are grateful to Elliott Sober, Adrian Erasmus, Simine Vazire, Stephen Crowley and Max Kasy, and audiences in several conferences and colloquia. We thank Zhaodong Chen for research assistance.

Appendix

Twelve hypotheses from the pre-analysis plans of Casey et al. (2012). We call the governance intervention X.

- H_1: "X creates functional development committees"
- H_2: "Participation in X improves the quality of local public services infrastructure"
- H_3: "Participation in X improves general economic welfare"
- H_4: "Participation in X increases collective action and contributions to local public goods"
- H_5: "X increases inclusion and participation in community planning and implementation, especially for poor and vulnerable groups; X norms spill over into other types of community decisions, making them more inclusive, transparent and accountable"
- H_6: "X changes local systems of authority, including the roles and public perception of traditional leaders (chiefs) versus elected local government"
- H_7: "Participation in X increases trust"
- H_8: "Participation in X builds and strengthens community groups and networks"
- H_9: "Participation in X increases access to information about local governance"
- H_{10}: "X increases public participation in local governance"
- H_{11}: "By increasing trust, X reduces crime and conflict in the community"
- H_{12}: "X changes political and social attitudes, making individuals more liberal toward women, more accepting of other ethnic groups and 'strangers,' and less tolerant of corruption and violence"

References

Backhouse R.E., Morgan M.S. (2000), "Introduction: is Data Mining a Methodological Problem?," *Journal of Economic Methodology* 7 (2): 171–181.

Barnes E.C. (2008), *The Paradox of Predictivism*, Cambridge University Press, Cambridge.

Bright L.K. (2017), "On Fraud," *Philosophical Studies* 174 (2): 291–310.

Brodeur A., Lé M., Sangnier M. et al. (2016), "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics* 8 (1): 1–32.

Camerer C.F., Dreber A., Holzmeister F. et al. (2018), "Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* Between 2010 and 2015," *Nature Human Behaviour* 2 (9): 637–644.

Casey K., Glennerster R., Miguel E. (2012), "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan," *The Quarterly Journal of Economics* 127 (4): 1755–1812.

Chambers C.D. (2013),"Registered Reports: A New Publishing Initiative at Cortex," *Cortex* 49 (3): 609–610.

Chambers C.D., Feredoes E., Muthukumaraswamy S.D. et al. (2014), "Instead of 'Playing the Game' it is Time to Change the Rules: Registered Reports at AIMS Neuroscience and Beyond," *AIMS Neuroscience* 1 (1): 4–17.

Christensen G., Miguel E. (2018), "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature* 56 (3): 920–980.

Coffman L.C., Niederle M. (2015), "Pre-Analysis Plans have Limited Upside, Especially where Replications are Feasible," *Journal of Economic Perspectives* 29 (3): 81–98.

Diaconis P., Mosteller F. (1989), "Methods for Studying Coincidences," *Journal of the American Statistical Association* 84 (408): 853–861.

Douglas H., Magnus P.D. (2013), "State of the Field: Why Novel Prediction Matters," *Studies in History and Philosophy of Science Part A* 44 (4): 580–589.

Dwan K., Altman D., Arnaiz J. et al. (2008), "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias," *PLoS ONE* 3 (8): e3081.

FDA (1997), "Food and Drug Administration Modernization Act," *105th U.S. Congress, U.S. House of Representative Bill*, URL = https://www.congress.gov/bill/105th-congress/senate-bill/830 [Accessed 13.07.2020].

Findley M.G., Jensen N.M., Malesky E.J. et al. (2016), "Can Results-Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study," *Comparative Political Studies* 49 (13): 1667–1703.

Foster A., Karlan D., Miguel E. (2018), "Registered Reports: Piloting a Pre-Results Review Process at the Journal of Development Economics," *World Bank Development Impact Blog*, URL = https://blogs.worldbank.org/impactevaluations/registered-reports-piloting-pre-results-review-process-journal-development-economics [Accessed 02.07.2020].

Frankel A., Kasy M. (2020), "Which Findings should be Published?," URL = https://maxkasy.github.io/home/files/papers/findings.pdf [Accessed 25.08.2020].

Frisch M. (2015), "Predictivism and Old Evidence: A Critical Look at Climate Model Tuning," *European Journal for Philosophy of Science* 5 (2): 171–190.

Glymour C. (1980), *Theory and Evidence*, Princeton University Press, Princeton, N.J.

Head M.L., Holman L., Lanfear R. et al. (2015), "The Extent and Consequences of P-Hacking in Science," *PLoS Biology* 13 (3): e1002106.

Howson C. (1991), "The 'Old Evidence' Problem," *The British Journal for the Philosophy of Science* 42 (4): 547–555.

Howson C., Franklin A. (1991), "Maher, Mendeleev and Bayesianism," *Philosophy of Science* 58 (4): 574–585.

Humphreys M., Sanchez De la Sierra R., Windt P.V.D. (2013), "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration," *Political Analysis* 21 (1): 1–20.

Ioannidis J.P.A. (2005), "Why Most Published Research Findings are False," *PLoS Medicine* 2 (8): e124.

Ioannidis J.P.A. (2008), "Why Most Discovered True Associations are Inflated," *Epidemiology* 19 (5): 640–648.

Leamer E.E. (1983), "Let's Take The Con out of Econometrics," *American Economic Review* 73 (1): 31–43.

Leonelli S. (2016), *Data-Centric Biology: A Philosophical Study*, University of Chicago Press, Chicago.

Libgober J. (Forthcoming), "False Positives and Transparency," *American Economic Journal: Microeconomics*, URL = https://www.aeaweb.org/articles?id=10.1257/mic.20190218 [Accessed 01.10.2020].

Maher P. (1988), "Prediction, Accommodation, and the Logic of Discovery," [in:] *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol. 1, A. Fine, J. Leplin (eds.), Philosophy of Science Association, East Lansing, MI: 273–285.

Mayo D.G. (1996), *Error and the Growth of Experimental Knowledge*, University of Chicago Press, Chicago.

Miguel E., Camerer C., Casey K. et al. (2014), "Promoting Transparency in Social Science Research," *Science* 343 (6166): 30–31.

Nosek B.A., Lakens D. (2014), "Registered Reports: A Method to Increase the Credibility of Published Results," *Social Psychology* 45 (3): 137–141.

Olken B.A. (2015), "Promises and Perils of Pre-Analysis Plans," *Journal of Economic Perspectives* 29 (3): 61–80.

Pagan A. (1987), "Three Econometric Methodologies: A Critical Appraisal," *Journal of Economic Surveys* 1 (1–2): 3–23.

Phillips P.C.B. (1988), "Reflections on Econometric Methodology," *Economic Record* 64 (4): 344–359.

Pearson T.A., Manolio T.A. (2008), "How to Interpret a Genome-Wide Association Study," *Journal of the American Medical Association* 299 (11): 1335–1344.

Simmons J.P., Nelson L.D., Simonsohn U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science* 22 (11): 1359–1366.

Sober E. (2015), *Ockham's Razors. A User's Manual*, Cambridge University Press, Cambridge.

White R. (2003), "The Epistemic Advantage of Prediction over Accommodation," *Mind* 112 (448): 653–683.

Worrall J. (2014), "Prediction and Accommodation Revisited," *Studies in History and Philosophy of Science Part A* 45 (1): 54–61.