HARKING: FROM MISDIAGNOSIS TO MISPRESCRIPTION

AYDIN MOHSENI

Abstract. The practice of HARKing—hypothesizing after results are known—is commonly maligned as undermining the reliability of scientific findings. There are several accounts in the literature as to why HARKing undermines the reliability of findings. We argue that none of these is right and that the correct account is a Bayesian one. HARKing can indeed decrease the reliability of scientific findings, but it can also increase it. Which effect HARKing produces depends on the difference of the prior odds of hypotheses characteristically selected ex ante and ex post to observing data. Further, we show how misdiagnosis of HARKing can lead to misprescription in the context of the replication crisis.

1. Introduction

In a 2019 article in *Nature*, the author, psychologist Dorothy Bishop, describes HARKing as one of "the four horsemen of the reproducibility apocalypse," along with publication bias, low statistical power, and p-hacking (Bishop 2019, p. 435). The practice of HARKing—hypothesizing after results are known—is commonly maligned as undermining the reliability of scientific findings.[1] There are several accounts in the literature as to why HARKing undermines the reliability of findings. Scholars have argued that HARKing undermines frequentist guarantees of long-run error control, that it violates a broadly Popperian picture of science, and misrepresents hypotheses formulated ex post to observing the data as those formulated ex ante. We argue that none of these accounts correctly identify *why* HARKing can undermine the reliability of findings, and that the correct account is a Bayesian one.

We will show that HARKing can indeed decrease the reliability of scientific findings, but that there are conditions under which HARKing can actually increase the reliability of findings. In both cases, the effect of HARKing on the reliability of findings is determined by the difference of the prior odds of hypotheses characteristically selected ex ante and ex post to observing data. To make this precise, given a natural model of null hypothesis significance testing, we provide necessary and sufficient conditions for HARKing to decrease the reliability of scientific findings.

[1] See, for example, Kerr (1998); John et al. (2012); Rubin (2017) and Murphy and Aguinis (2019).

The aim of this paper is not to defend the practice of HARKing. Insofar as HARKing involves disclosing less than complete information to those who wish to *learn and act* based on scientific findings, it is clearly epistemically undesirable.[2] HARKing can also be ethically and pedagogically undesirable, insofar as it involves intentional deception or presenting an inaccurate model of science to students. Rather, the aim here is to clarify the relationship between HARKing and the reliability of scientific findings.

Understanding HARKing is important on at least two counts. Historically, HARKing is closely tied to questions regarding the relationship between prediction and accommodation. These questions have engaged philosophers at least as early as Mill (Mill 1843), were made central in the philosophy of science by Popper (Popper 1934), and continue to be of concern in contemporary discussions in scientific epistemology (Hitchcock and Sober 2004). As mentioned, HARKing is also imputed to be among the questionable research practices contributing to the crisis of replication in the social and biological sciences, which has rightly become a subject of interest to philosophers of science.[3] A better understanding of HARKing sheds light on both these issues.

My strategy for demonstrating that standard accounts for why HARKing leads to unreliable findings are incorrect is as follows. Each account claims that HARKing undermines the reliability of scientific findings because HARKing exhibits a particular property φ. We show that HARKing can increase the reliability of findings while still satisfying property φ and, hence, φ cannot explain why HARKing is in fact bad for the reliability of findings. Instead, we provide a Bayesian analysis of HARKing that provides necessary and sufficient conditions for when HARKing worsens or improves the reliability of findings.

In §2, we summarize several accounts of why HARKing is bad for the reliability of scientific findings. In §3, we present clear criteria for the reliability of scientific findings with which to measure the effect of HARKing relative to specific alternatives. In §4, we present a standard model of hypothesis testing with which to reason about the statistical consequences of HARKing. In §5, we provide necessary and sufficient conditions for when HARKing improves and worsens the reliability of findings. In §6, we show how misdiagnosis of HARKing ramifies into misguided proposals for redefining statistical significance in the context of the replication crisis. In §7, we conclude with a discussion.

[2] For precise, decision theoretic formulations of this observation see the value of knowledge theorems of Savage (Savage 1954), Good (Good 1967), and Skyrms (Skyrms 1990, Ch. 4).

[3] For excellent philosophical examinations of social and epistemic issues involved in the replication crisis see Romero (2019, 2020); Romero and Sprenger (2020); Heesen (2018); Bruner and Holman (2019); Bright (2017); Bird (2018); and Machery (2020).

2. HARK! Who Goes There?

First, let us be clear about what we mean here by HARKing. The term 'HARKing' was first coined by social psychologist, Norbert Kerr, in his 1998 article "HARKing: hypothesizing after results are known." Kerr defines HARKing as "...presenting a post hoc hypothesis in the ... [study] report as if it were an a priori hypothesis" (Kerr 1998, p. 197).[4] HARKing occurs when a researcher selects her study hypothesis after observing the data in and reports this hypothesis as if it had been formulated prior to observing the data—that is, as if it had been a prediction. This is typically contrasted with the normative protocol in which the researcher selects a hypothesis prior to observing the data, and then, after observing her data, reports whether the hypothesis attained significance given some conventional threshold for statistical significance.

In his 1998 article, Kerr actually anticipates many of the now standard objections to the practice of HARKing. These include taking unjustified statistical license, propounding theories that cannot pass Popper's falsifiability criterion, and disguising post hoc explanations as a priori explanations (Kerr 1998, p. 211). Since then, and especially in light of the replication crisis, philosophers of science and scholars in the social and biomedical sciences have elaborated and propounded these accounts (Rubin 2017).

Several variants of HARKing exist in the literature and should be distinguished (Rubin 2017). STARKing, or story telling after results are known, is where a finding is presented along with a narrative produced ex post to observing the data meant to bolster the plausibility of that finding. THARKing, or transparent hypothesizing after results are known, is where it is clearly communicated that the study hypothesis was selected after the data were observed (Hollenbeck and Wright 2016). We concern ourselves here only with HARKing. In particular, we are concerned here with accounts of the epistemic effect of HARKing: why, precisely, it undermines the *reliability* of scientific findings, as presented in (Kerr 1998) and in other influential accounts such as (Rubin 2017) and (Mayo 2019).

2.1. **HARKing as Undermining Error Control.**

The first account of the epistemic problem of HARKing emerges straightforwardly from classical, frequentist philosophy of statistics that concerns itself with error-control. In the context of hypothesis testing, a central strand of frequentist thought locates the reliability of tests in terms of their guarantees of controlling the long run frequencies of Type I and Type II error in hypothetical repetitions of those tests (Lehmann 1993).

[4] Kerr employs 'a priori' and 'a posteriori' to mean before and after the event of observing one's study data. We use the terms 'ex ante' and 'ex post' for these to avoid confusion with the standard philosophical meanings of the former terms.

An example of such a guarantee is as follows. Consider a hypothesis test with a conventional significance threshold of $\alpha \in [0, 1]$ (corresponding to its Type I error rate). A sample of data is collected for which a test statistic, t, is determined.[5] A decision to reject or fail to reject the hypothesis is made as follows. On the assumption that the null hypothesis is true,[6] one determines the p-value for the test, or the probability of having observed a test statistic at least as extreme was actually observed, $p = P(T > t|H_0)$. If this value meets the significance threshold, $p < \alpha$, then the null hypothesis is rejected. If the threshold is not met, one fails to reject the null. In a world where the null hypothesis is true, such a test produces mistaken rejections of the null hypothesis $100 \times \alpha$ percent of the time if the test were repeated infinitely many times.[7]

HARKing undermines such guarantees. When a researcher engages in HARKing, she waits until after she observes her data and then selects a hypothesis to report from among those that are statistically significant. To drive our point home, consider a researcher who has infinitely many probabilistically independent hypotheses from which she may choose. Further, imagine that all of her hypotheses are false. For any positive Type I error rate, $\alpha > 0$, she will obtain statistically significant results, as mistaken rejections of the null are now a certainty. If she engages in HARKing, she will only ever report significant findings, even though all of her candidate hypotheses are false, and so guarantees of error-control of the sort just described become ill-defined.

Consider the following formulation of the problem by Rubin: "For example, if a researcher tests 20 hypotheses with an alpha level of .05, then he has a 64.15% chance of making at least one Type I error. However, if his results confirm only one of these hypotheses, and he decides to suppress the other 19 disconfirmed hypotheses, then he will give the incorrect impression that he only conducted a single hypothesis test and that, consequently, he only had a 5% chance of making a Type I error." (Rubin 2017, p. 14) Bishop echoes a familiar refrain in describing the consequences of HARKing: "P-values are meaningless when taken out of context of all the analyses performed to get them." (Bishop 2019, p. 435)

This is indeed correct; HARKing undermines frequentist guarantees of long run error control. However, we are interested in the *reliability* of scientific findings, and the types of error that the frequentist promises to control—i.e., Type I and Type II error rates—simply do not capture the reliability of findings. The Type I and II error rate of a test tell us that *if* the hypothesis is true or false, *then* what

[5] For example, the test statistic may be the mean of the difference or association between two variables in a data set.

[6] And also that the inductive assumptions of the test—e.g., normality, homoscedasticity, and so on—hold true.

[7] More precisely, this occurs *almost surely* with respect to the measure over the infinite sequence of outcomes.

is the long run frequency of errors to be expected.[8] We discuss this further in §3 where we provide a natural and practicable measure of the reliability of findings.

2.2. **HARKing as Failing to Provide a Severe Test.** The second account maligns HARKing for violating a broadly Popperian picture of science. The basic idea is that when a hypothesis is selected ex post to observing data for its compatibility with those data, then it could not be reliably disconfirmed by those data. This is described by Rubin as the objection that HARKing is "problematic for scientific progress because it results in hypotheses that are always confirmed and never falsified by the results." (Rubin 2017, p. 2) Kerr states this plainly: "HARKed hypotheses fail Popper's criterion of disconfirmability" (Kerr 1998, 205).

That said, we need to sharpen this objection, as *deductive* falsification is typically not feasible for statistical tests of hypotheses. The most sophisticated version of this view which has risen to prominence in the the philosophy of statistics is the severe testing account formulated by Deborah Mayo (Mayo 2019). On this account, a statistical test provides us with corroboration of a hypothesis insofar as it submits that hypothesis to a severe test, where a severe test is one that would reliably detect an error in that hypothesis if one were present (Mayo 2019). This is Popper's criterion of falsifiability adapted to the statistical context.

It is clear why a hypothesis reported via HARKing fails to satisfy the requirements of a severe test. Recall that under HARKing the researcher selects her hypothesis after observing from the set of hypotheses that are significant given her data. The reported results of her study are significant by construction. She could have failed to report the hypothesis, but she would have not have reported its 'falsification' by the data if it were not significant. Thus, HARKing fails to provide warrant for belief in hypotheses because it fails to expose them to opportunities for statistical 'falsification'.

As with frequentist error rates, the relationship between the severity of tests and the reliability of scientific findings is not a direct one. What we will see is that HARKing can fail both to provide well-defined frequentist error rates or meet the requirements of severe testing, and yet produce more reliable findings, and so, whatever virtues these accounts capture, they fails to explain the interaction of HARKing with the reliability of scientific findings.

2.3. **HARKing as Misrepresentation.** Our third account, from (Kerr 1998), posits that HARKing misrepresents hypotheses formulated ex post to observing the data as those formulated ex ante. As mentioned, we have no truck with the version of this objection that locates the ill of HARKing in ethical terms. But

[8] And the p-value tells us the probability, conditional on the null being true, of observing data at least as extreme as was actually observed.

this objection rarely goes beyond observation of the fact of misrepresentation to identify why, precisely, misrepresentation of this sort should negatively effect the reliability of findings. That it does so appears to be just assumed.

My analysis will explain the relationship between HARKing and the reliability of findings. In particular, we will show that the last objection is closest to the mark: it is misrepresentation of the hypotheses that can be epistemically detrimental. When HARKing is bad, it is because it can lead us to mistakenly infer that the hypothesis reported via HARKing enjoyed greater ex ante evidential support than it in fact did, and so garners greater ex post confidence than it in fact deserves. Let us to turn to our Bayesian account.

3. The Reliability of Scientific Findings

Our argument requires an adequate specification of the reliability of scientific findings with which to determine whether a given account of HARKing correctly diagnoses the effect of HARKing on the reliability of findings. For this, we draw on existing statistical concepts, specifically false discovery and false omission rates, to formalize a natural notion of reliability in the context of null hypothesis significance testing. I argue for why this is a more apt characterization of reliability than frequentist guarantees of error-control, specifically, Type I and II error rates, or data-dependent notions such as the 'severity' of a test.

Classically, we want our epistemic methods to produce fewer false beliefs and more true ones. In the statistical context, we can ask that fewer of the results we declare significant be false and more of them be true, and fewer of the results we declare non-significant be true. These correspond to the requirement that our methods exhibit lower rates of false discovery and false omission respectively.[9]

The *false discovery rate* (FDR) of a population of studies is the expected proportion of its findings (rejections of the null hypothesis) that are false findings (where the null hypothesis is true). For a given method of reporting findings M and population of studies the FDR corresponds to:

$$FDR(M) = Pr(H_0 \mid \text{significant}; M) = \frac{Pr(H_0, \text{significant}; M)}{Pr(\text{significant}; M)}.$$[10]

That is, the ratio of false and significant findings over over all significant findings. Intuitively, the reliability of research increases as the false discovery rate

[9] The notion of false discovery rate of studies was first introduced to the statistical sciences by Bejamin and Hochberg (Benjamini and Hochberg 1995) to understanding the expected frequencies of true and false hypotheses in the context of multiple testing.

[10] Note that the probability of study outcomes is employed here as it is mathematically equivalent to the corresponding expected fraction of study outcomes. For example, the probability of a given study hypothesis obtaining significance while the null is true is equivalent to the expected fraction of study hypotheses that obtain signficance while their nulls are true.

decreases. Indeed, in the context of the replication crisis, the rate at which findings in a literature fail to replicate under more stringent tests is an estimator for the false discovery rate of that literature.

The *false omission rate* (FOR) of a literature is the expected proportion of its negative findings (failures to reject the null hypothesis) that are false negative findings (where the alternative hypothesis is true). For a given method of reporting findings M and population of studies the false omission rate corresponds to:

$$FOR(M) = Pr(H_1 \mid \text{not significant}; M) = \frac{Pr(H_1, \text{not significant}; M)}{Pr(\text{not significant}; M)}.$$

That is, the ratio of true and non-significant findings over all non-significant findings. Intuitively, the reliability of studies increases whenever their false omission rate decreases. In particular, FOR provides a measure of how poorly research detects the truth of hypotheses.

Protocols for selecting hypotheses and reporting results can affect both the false discovery and omission rates of a population of studies. For our analysis, we consider broadly two sorts of protocols: one which requires the researcher to commit to a hypothesis prior to observing the data, and another that allows her to select her hypothesis only after observing the data. Some changes in methods increase one type of error while decreasing the other.[11]

Definition (Reliability of a Method)**.** Let a population of studies be specified by the significance threshold $\alpha \in (0, 1)$ and mean power[12] $\beta \in (0, 1)$ of its tests along with the prevalence of true hypothesis $\pi \in (0, 1)$ among the set of hypotheses selected for testing.

Say that a given reporting method M is *more reliable* with respect to *false discovery* than another M' if, for any such population of studies, findings produced under M yield a lower false discovery rate than those under M'.

$$FDR(M) \leq FDR(M'), \quad \text{for any } \pi, \alpha, \beta \in (0, 1).$$

Similarly, say that M is *more reliable* with respect to *false omission* than M' if, for any such population of studies, findings produced under M produce a lower false omission rate than those under M'.

$$FOR(M) \leq FOR(M'), \quad \text{for any } \pi, \alpha, \beta \in (0, 1).$$

[11] For example, it is well-known that lowering the threshold for statistical significance will tend to decrease false discovery while increasing false omission.

[12] Statistical power is the complement of Type II error. That is, the power of a test is the probability, conditional on the alternative hypothesis being true, of correctly rejecting the null hypothesis.

Note that we require that the studies be nontrivial in that there is some non-zero fraction of true and false hypotheses to be tested, that is $\pi \in (0, 1)$, otherwise improvements in method could not improve the FDR and FOR of studies.[13]

These provides natural, practical measures of reliability. But what of classical Type I and II error rates? To see why Type I and II error rates cannot capture the reliability of findings, consider a domain of scientific study in which no true hypotheses are available—that is, where the null hypothesis is always true. Regardless of the Type I and II error rates of tests, all significant findings will be false findings. More generally, one and the same frequentist guarantee—Type I and II error rates—is compatible with any reliability of scientific findings since reliability is critically a function of the prevalence of true hypotheses put to test.

Other measures of reliability might be sought in data-dependent measures, such as p-values or the severity of a test.[14] Yet the same shortcoming applies to such measures. Insofar as a measure ignores the prevalence of true hypotheses submitted to testing, it will assign the same measure of reliability to a set of findings assured to be false as to a set of findings assured to be true.

In the context of classical null hypothesis significance testing, what we may want from a measure of reliability of scientific findings is that more of those claimed to be statistically significant are in fact true and more of those claimed not to be significant are in fact false.[15] The false discovery and omission rates of hypotheses captured just this; and with them in hand, we can proceed to an analysis of the interaction of HARKing and the reliability of scientific findings.

It might be argued that the prevalence of true hypotheses is not a quantity generally known to us, especially given the distortions produced by publication bias, so a measure of reliability that uses this unknown quantity is useless. There are two problems with this objection. First, reasonable estimates of the prevalence of true hypotheses in a domain are difficult but not impossible to produce. Indeed, there is work on this topic (Dreber et al. 2015). Second, and more to the point, we can still entertain the hypothetical: we can ask what the false discovery and omission rates of different reporting protocols *would be* given different underlying prevalences of true hypotheses submitted for testing. And we can learn if one

[13] Though improvements of methods could improve one of them. For example, if all candidate hypotheses are true, a method could not improve the FDR (since there could be no false discoveries), but it could improve the FOR merely by assigning significance to more results.

[14] The severity of a test can be thought of as data-dependent analogue of statistical power (Mayo-Wilson and Fletcher 2019).

[15] Alternatively, one may wish to move to leave the NHST paradigm for, for example, a fully Bayesian approach to analyzing scientific findings in which one applies credences over the truth of one's study hypotheses. The authors cautiously endorse proposals of this sort (see, for example, (Etz and Vandekerckhove 2016).), but recognize that it is worthwhile improving existing statistical practices even as we work toward more substantive, long term changes in method.

method outperforms another *regardless* of whether truth is a rare disease or as common as pig tracks.

4. When HARKing Cannot be Bad

Consider a world in which all study hypotheses have equal prior odds. We will show that in such a world HARKing cannot lessen the reliability of findings. Let us see why this is so and consider the implications of this fact.

First, a note on how to interpret the prior odds of hypotheses. In our model, the prior odds of a hypothesis are to be understood in terms of a well-defined prevalence of true study hypotheses. A study hypothesis H_i belongs to a set of candidate hypotheses, $\boldsymbol{H} = \{H_i\}_{j=1}^{n}$, from which it is selected by the researcher. A given fraction of the hypotheses in the set, $\pi \in (0, 1)$, are true and their complement false. If hypotheses are randomly selected from this set for testing, the prior probability of a study hypothesis is just the probability of selecting a true hypothesis from this set $Pr(H_i) = \pi$.[16][17]

Now, let us define our research methods. In the endorsed picture of hypothesis testing, the researcher selects her hypothesis, H_i, from the set of possible hypotheses prior to observing her data. Only then does she observe her data, and then she reports whether her predicted hypothesis, H_i, was statistically significant given the conventional threshold for significance, α. Call this protocol *prediction* and denote it M^p.

In contrast, under a protocol of *HARKing* the researcher first observes her data, and then selects a hypothesis H_i at random from the set of hypotheses that have turned out to be statistically significant in light of her data $\{H_i \in \boldsymbol{H} | p_i < \alpha\}$, if the set is nonempty. Denote this protocol M^h.

We summarize the preceding two reporting protocols as follows.

(1) *Prediction* M^p: Prior to observing the data, select a single hypothesis to test. Report the hypothesis if turns out to be significant.
(2) *HARKing* M^h: After observing the data, randomly select a hypothesis from among those that are significant (if there are any) and report the hypothesis.

Note that we assume that a hypothesis is reported only if it is significant. This reflects the reality of publication bias and the concomitant file drawer effect. That said, nothing critical turns on this assumption, and later we will allow for some

[16] Equivalently, the prior odds of the hypothesis will be $1 : (\pi^{-1} - 1)$. We use the terms 'prior odds' and 'prior probabilities' to denote the same quantity.

[17] For the Bayesian, the analysis is more straightforward: the prevalence of true hypotheses is just her prior. The stipulation of a process of random selection of hypotheses is provided to make the analysis more palatable to an interlocutor skeptical of ostensible subjectivity of the Bayesian approach; priors here correspond to objective elements of the probability model—fractions of true hypotheses in a well-defined population of hypotheses—and not subjective beliefs.

fraction of statistically non-significant findings to be reported as well when we consider the false omission rates of protocols.

Given our specification of these protocols, we can demonstrate the following. (All proofs are provided in the appendix.)

Proposition 1. When hypotheses are selected from the same set of candidate hypotheses with fraction $\pi \in (0, 1)$ true hypotheses, then prediction yields the same false discovery rate as HARKing.

$$FDR(M^p) = FDR(M^h)$$

That is, in such a case, HARKing is as reliable with respect to false discovery as prediction. The logic of the result is simple. By stipulating that hypotheses be selected at random, we have set the fraction of true hypotheses selected via both prediction and via HARKing to be equal. And, in both cases, selected hypotheses are reported only if they are significant. Thus, the fraction of hypotheses that are true as well as significant is identical. See figure 1, where the only functional difference between the protocols is that prediction checks only a single hypothesis at a time, whereas HARKing checks all candidate hypotheses; both methods produce identical statistics in reported hypotheses. The only difference is that a researcher employing HARKing filters a larger set of hypotheses for statistical significance. In this world, the researcher employing harking is simply more efficient in filtering hypotheses for significance—while the fraction of her true and false discoveries is the same, her absolute rate of discovery is strictly greater.

What of false omission rates? The worries regarding HARKing are typically focused on its contribution to high false discovery rates, which correspond to low replication rates. Missing out on true hypotheses is not typically voiced as a concern, especially because whatever is presumed to make us more likely to declare both true and false results as significant is obviously more likely to make true results significant. But this claim is also false of HARKing in the world where hypotheses are selected for filtration via prediction and HARKing in a way that gives them equal prior odds. This fact follows the same reasoning as the equality of false discovery rates (and is proved in the mathematical appendix). Both prediction and HARKing protocols filter hypotheses with the same frequencies, only HARKing does so more efficiently.

Let us revisit an assumption. When engaging in HARKing and confronted with multiple significant hypotheses, we assumed that the researcher selects one at random. What if, instead, she reports the most impressive, publication-worthy result—the result of the lowest p-value? This natural description yields the following reporting protocol.

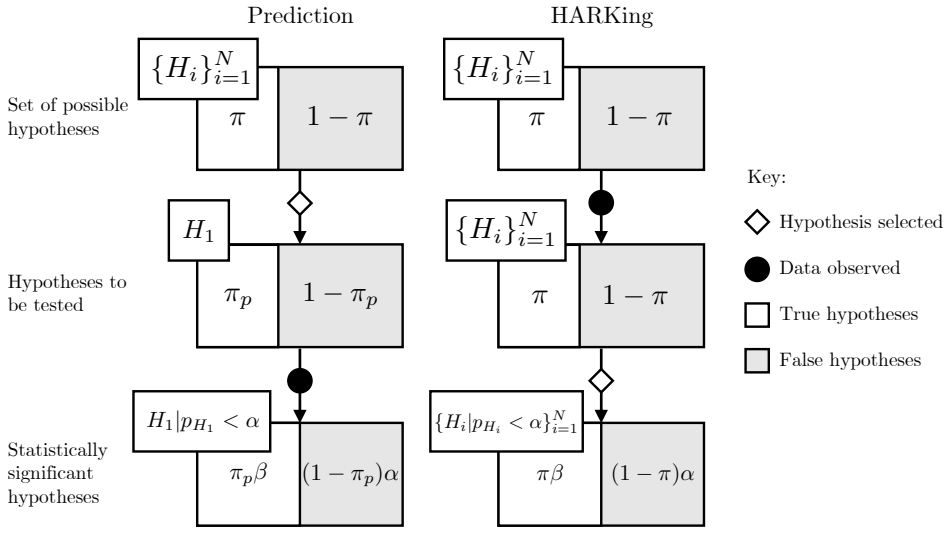

FIGURE 1. The filtration of study hypotheses for reporting via prediction and of HARKing.

(3) *Selective HARKing* M^{sh}: After observing the data, select the hypothesis with the lowest p-value from among those that are significant (if there are any) and report the hypothesis.

Hypotheses that yield lower p-values are, on average, more likely to be true. Thus, the population of studies reported via selective HARKing will be composed of more true findings than either of those produced via prediction or HARKing. And as more possible hypotheses are considered, the lower the expected p-value of the hypothesis with the least p-value will be, and the greater the fraction of true reported hypotheses.

Proposition 2. When hypotheses are selected from the same set of possible hypotheses with fraction $\pi \in (0, 1)$ true hypotheses, and there are more than two candidate hypotheses, then:

(1) Prediction yields a false discovery rate exceeding that of selective HARKing, $FDR(M^p) > FDR(M^{sh})$.
(2) The false discovery rate for selective HARKing is decreasing in the size of the set of candidate hypotheses L.

That is, a slightly more sophisticated version of HARKing can produce strictly and substantially more reliable findings than prediction.

5. When HARKing Must be Bad

When is HARKing bad for the reliability of scientific findings? Under the conditions we described—equal prior odds of hypotheses selected via prediction and HARKing—selecting hypotheses ex post to observing the data cannot undermine the reliability of one's findings. Thus, for HARKing to be bad in the course of normal scientific practice, one of the assumptions of our model must not obtain. The natural candidate is that scientists do not in fact choose their hypotheses at random.

A researcher's choice of study hypothesis tends to be informed by her domain knowledge. The hypotheses she chooses prior to observing the data may be informed by theory, previous findings in the literature, and common sense. We can imagine that, when a researcher is suitably informed, a hypothesis she selects is more likely to be true than a hypothesis selected at random from the set of hypotheses that are merely logically consistent with her data. In such a case, we can expect the researcher to do better than chance prior to observing the data. Under prediction, such hypotheses are then filtered by their statistical significance after the data have been observed.

When a researcher is engaged in HARKing, however, she is no longer filtering the set of possible hypotheses via her domain knowledge. Rather, the full set of hypotheses consistent with her data are submitted to the filter of statistical significance.

Formally, this corresponds to a model in which the researcher chooses from two sets of hypotheses with different fractions of true and false hypotheses. When she chooses ex ante to observing her data, via prediction, she tests a subset of hypothesis with some prior odds; and when she chooses ex post to observing her data, via HARKing, she chooses from a superset with different prior odds. For HARKing to be bad in such a world, it *must* be the case that the prevalence of true hypotheses is greater in the set of hypotheses that are selected for prediction than in the set of all possible, candidate hypotheses. This is captured in the following proposition.

Proposition 3. Let hypotheses selected via *prediction* and *HARKing* exhibit base rates π_p and π, respectively. Then prediction is more reliable than HARKing with respect to false discovery, $FDR(M^p) < FDR(M^h)$, just in case $\pi_p > \pi$.

This explains why misrepresentation of hypotheses selected ex post as those selected ex ante to observing data can be pernicious. If we expect a researcher to have meaningful domain knowledge, then we should expect her choice of hypothesis to be informative and so for the hypothesis she selects ex ante to be more likely to be true. If we are misled on this count, and the hypothesis reported does

not boast such support, then we will (intuitively) assign too great a credence to her findings.

Proposition 3 also tells us when HARKing will produce more reliable findings than prediction. Consider an unlucky world in which the scientist's judgment is anti-correlated with the truth. That is, when the researcher makes predictions, she selects false hypotheses at a rate that is worse than chance. In this world, it is better to HARK than to predict.

Such an unfortunate world is not just a modeling fantasy. It is well-documented that in the domains of social and political punditry, humans can perform the impressive feat of doing consistently worse than chance.[18] More generally, prediction may fare unfavorably when we are confronted with problems where our domain knowledge is limited—as in cases with limited theory, few or no prior studies, or where common sense is largely unhelpful. One can think of analyses of any complex system where it is to be expected that a multitude of factors conspire to produce effects of interest. In such cases, a small increase in the false discoveries produced by ex post hypothesis selection may be compensated for by a greater decrease in false omission rates that may redound to leads for future, confirmatory research.

In sum, prediction can yield greater false discovery rates than HARKing, or HARKing can produce greater false discovery; what determines which obtains is the prior odds of hypotheses submitted to each. In the real world, we can expect that the prior odds of hypotheses submitted for HARKing is determined by the challenge of the domain, and the difference of prior odds of hypotheses submitted for prediction is determined by researcher judgment.

6. Jumping the HARK: From Misdiagnosis to Misprescription

Misdiagnosis of HARKing can ramify in the misprescription of solutions to the replication crisis. One prominent line of thinking in the literature is that if questionable research practices such as HARKing are bad because they make studies more likely to yield false positive results, then one natural solution is to lower the conventional threshold for statistical significance to compensate. A recent statement signed by over 50 prominent methodologists proposes redefining statistical significance in just this way (Benjamin et al. 2018). Stated plainly,

> "For fields where the threshold for defining statistical significance for new discoveries is $p < 0.05$, we propose a change to $p < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields." (p. 6)

[18] Cf. Tetlock and excellent work in (Tetlock 2017) for a presentation of the literature on expert political judgment.

This is not a prima facie unreasonable proposal. Fields such as genomics and high energy physics have profitably set more stringent standards for their conventional significance threshold in the context of gene-wide association studies (Franklin 2013).[19] But there are differences. John Ioannidis expresses worry regarding the efficacy of lowering the significance threshold in the social and biomedical sciences, citing the relatively greater researcher degrees of freedom in those disciplines, "Adopting lower p-value thresholds may... [produce] collateral harms... bias may escalate rather than decrease if researchers... try to find ways to make the results have lower p-values" (Ioannidis 2018, p. 1430). The model we present can be seen as demonstrating a precise realization of Ioannidis' worry.

To see why such an intervention can been seen a solution to the ills of questionable research practices such as HARKing, consider the following simple model where researchers are engaging in a strategic mixture of both prediction and HARKing protocols: she follows a protocol of prediction when she can, and a protocol of HARKing when she must in order to attain statistical significance for some findings, and so to publish her study. Call this method *fallback Harking* M^{fh}. This can be thought of as a plausible approximation of what many researchers in certain domains in fact do (John et al. 2012).

Here, a study consists of procuring data against which a set of N logically independent hypotheses $\{H_i\}_{i=1}^{N}$ may be tested. In fallback HARKING, prior to observing the data, a researcher selects the hypothesis for testing, H_1, that she judges is most likely true. Upon observing the data, if she finds that her hypothesis, H_1, is statistically significant then she reports it. If, on the other hand, she finds that her hypothesis is not statistically significant, she casts a broader net and turns to the $N-1$ other possible hypotheses $\{H_i\}_{i=2}^{N}$ and reports one that is significant, if such a one exists.[20]

Importantly, the researcher's prediction here is informed by her domain knowledge. The hypothesis she chooses prior to observing the data, H_1, is supported by some combination of theory, previous results, common sense, and so on, and so the hypotheses she chooses are on average more likely to be true than a random member of the other $N-1$ hypotheses, $P(H_1) = \pi_1 > \mathbb{E}[\{\pi_i\}_{i=2}^{N}]$. These other hypothesis may be true, but they are not, on average, as well supported by her domain knowledge.

[19] Though, for criticisms of the five-sigma rule, see (Lyons 2015) and (Lyons 2013).

[20] As with HARKing, a hypothesis is chosen at random from among the set of significant hypotheses. One could instead consider the case where the researcher chooses the significant hypothesis with the greatest hypothesis. The qualitative outcome of the model—the possibility of an increase of false discovery rate as the significance threshold is lowered—would not change as long as researcher judgment was sufficiently informed.

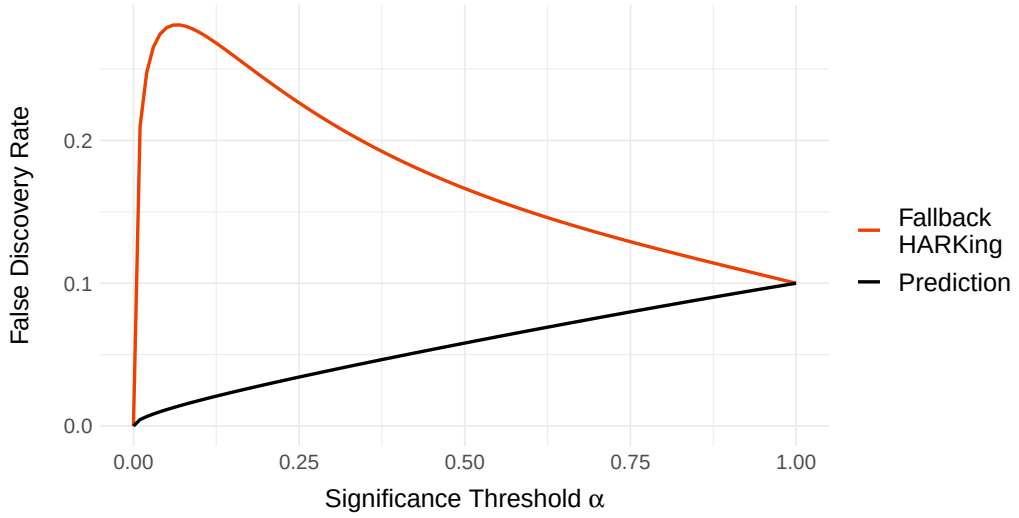

FIGURE 2. The false discovery rates for 'fallback HARKing' and 'prediction' protocols as functions of the significance threshold α when $\pi_p = 0.9, \pi = 0.1$, and $\beta = 0.2$.

Consider the effects of lowering the conventional threshold for statistical significance, α, in such a case—where researchers are engaged in a reporting protocol like fallback HARKing. Figure 2 shows the relationship between the significance threshold, α, and false discovery, FDR, in such a case. If all researchers adhere to the method of prediction, M^p, then lowering α expectably reduces the false discovery rate $FDR(M^p(\alpha))$ (see the black line in Figure 2). Similarly, if all researchers adhere to a protocol of HARKing, then false discovery rate decreases as α decreases.[21]

However, if researchers follow a protocol like fallback HARKing $M^{fb}(\alpha)$, the false discovery rate can actually increase (see the red line in Figure 2). The reason for this is that, as α decreases—and the evidential standard for significance becomes more stringent—researchers are less likely to attain statistical significance for their primary hypotheses, and so they are more likely to turn to scouring auxiliary hypotheses. That is, researchers must turn away from the few hypotheses with greater prior odds and toward the many hypotheses that exhibit potentially far lower prior odds. And since there are many more of the latter than the former, they are likely to find some that have attained significance by chance. These statistically significant 'fallback hypotheses', in turn, are more likely to be false discoveries.

[21] Though, of course, by the same token, lowering α must increase the false omission rat—a greater fraction of cases where the alternative hypothesis is true must fail to attain significance.

In short, when researcher judgment on a problem is well-informed, lowering the significance threshold can actually push researchers off of their fewer more promising hypotheses and onto dredging their many unpromising ones, and so can increase the false discovery rate of a population of studies.

When should we expect that lowering the significance threshold will increase false discovery? This is an empirical question with deep implications; and the answer will depend on several key factors. In particular, it will depend on the prevalence of behaviors approximating fallback HARKing,[22] the current value of the significance threshold and the extent to which it is lowered, the average power of the population of studies under question,[23] the prevalence of true hypotheses in the domain in question,[24] and, as our analysis reveals, the strength of researcher judgment.[25] Exploration of optimal values for statistical significance for a given domain given the distinctive methods and challenges of that domain must remain for future work.

Note, however, that correctly identifying the mechanism by which the HARKing affects the reliability of studies—differences in expected prior odds of hypotheses selected ex ante and ex post observing data—was crucial for identifying the *very possibility* of the undesirable consequences of redefining statistical significance just discussed. The effect could not be characterized by merely looking at either the Type I or II error rates of protocols, or by examining their p-values or test severity. One must attend to the reporting protocols with the prior odds of hypotheses each characteristically submits to tests.

7. HARKening Back to the Good Old Bayes'

We have seen that the standard accounts in the literature fail to explain why, precisely, HARKing undermines the reliability of scientific findings. The properties they claim account for why HARKing is bad obtain even when HARKing improves the reliability of findings. A Bayesian analysis elucidates the relationship between HARKing and the reliability of scientific findings: HARKing can increase or decrease the reliability of findings relative to prediction as a function

[22] For studies of the prevalence of questionable research practices, including HARKing, see (John et al. 2012) and (Head et al. 2015).

[23] For estimates of the average power of studies in psychology see (Szucs and Ioannidis 2017) and (Wassertheil and Cohen 2006).

[24] For recent work on this count in replication markets, see (Pawel and Held 2020)

[25] There is little work isolating the reliability of researcher judgment in hypothesis selection. However, recent work on replication markets shows that researchers predict the replication of the studies of their colleagues with remarkable accuracy (Dreber et al. 2015). Though, of course, it is possible that this good judgment extends to the hypotheses of others, and not to one's own. Further, such findings may produce adequate estimates for the replicability of studies, if not the pre-test prevalence of true hypotheses; though this suggest natural methods to estimate the latter using the former.

of the differences in the prior odds of hypotheses characteristically selected ex ante and ex post to observing data.

Further, we have conjectured that the natural mechanism for producing the difference in prior odds of hypotheses selected in prediction is researcher judgment. When a scientist is meaningfully informed in her ex ante choice of hypothesis then her prediction is formally equivalent to restricting the set of reported study hypotheses to a subset which will, on average, be more likely to be true. HARKing, on the other hand, is formally equivalent to failing to make such a restriction. Thus, when scientists are uninformed, or worse, systematically biased, prediction can correspond to restriction to a subset of hypotheses with lower expected prior odds, and so HARKing, or a plausible variant of HARKing, can outperform prediction in terms of the false discovery and omission rates of the findings produced.[26]

Moreover, it is important to bear in mind that the decrease in the false discovery rate produced by prediction comes at the cost of increase in false omission rates. Any non-trivial restriction of the set of candidate hypotheses must lower the absolute rate of discovery. This suggests that a more ethical version of HARKing, such as transparent HARKing, may be preferable in contexts where lowering the false omission rate matters more to researchers or policymakers than lowering false discovery rates.[27]

We have also argued that the misunderstanding of HARKing has consequences for proposed solutions to the replication crisis. In particular, we considered a recent proposal to redefine the conventional threshold for statistical significance and how such a proposal can lead to undesirable consequences in light of an accurate understanding of how HARKing affects the reliability of findings.

Our moral is that current accounts of HARKing that stem from a frequentist philosophy of statistics fail to explain the actual logic of its interaction with the reliability of scientific findings, that their misdiagnosis ramifies into misprescription for solutions in the context of the replication crisis, and that a Bayesian analysis of the problem makes this all clear.

[26] This provides a recommendation for when selecting hypotheses ex post is unequivocally likely to be better: in inference tasks where the set of plausible hypotheses tends to be very large, highly complicated, or where researchers are known to be biased.

[27] One can think of the characteristic differences in such preferences in the concrete examples of preliminary, exploratory analyses aimed at identifying promising future cancer treatments in contrast to confirmatory analyses of phase III clinical trials aimed at vetting whether a pharmaceutical should hit the shelves and be made available to the public.

Mathematical Appendix

For the following proofs it is assumed that we have non-extremal values for each the significance threshold, $\alpha \in (0,1)$, and mean power, $\beta \in (0,1)$, of tests as well as for the prevalence of true hypothesis, $\pi \in (0,1)$, in the set of possible hypotheses.

Proof of Proposition 1. Consider the false discovery rate for the prediction protocol M^p. Recall that, in prediction, a hypothesis is selected for testing prior to observing the data, and that after observing the data the selected hypothesis is reported only if it attains statistical significance.

Let $\pi = Pr(H_i)$ be the fraction of true hypotheses in the set of possible hypotheses; $\pi_p = Pr(H_1)$ the fraction of true hypotheses in the subset of hypotheses selected via prediction (first we consider the case where a hypothesis is selected at random for prediction; and so $\pi_p = \pi$); let $p = Pr(T > t|H_0)$ denote the p-value of the test; $\alpha = Pr(p \leq \alpha|H_0)$ the significance threshold of the test; and $1-\beta = Pr(p \leq \alpha|H_1)$ the power of the test. The false discovery rate for prediction is then obtained via Bayes rule.

$$\begin{aligned} FDR(M^h) &= \frac{Pr(p \leq \alpha, H_0)}{Pr(p \leq \alpha)} \\ &= \frac{Pr(p \leq \alpha|H_0)Pr(H_0)}{Pr(p \leq \alpha|H_0)Pr(H_0) + Pr(p \leq \alpha|H_1)Pr(H_1)} \\ &= \frac{\alpha(1-\pi_p)}{\alpha(1-\pi_p) + (1-\beta)\pi_p} \\ &= \Big(1 + \frac{\pi_p}{1-\pi_p}\frac{1-\beta}{\alpha}\Big)^{-1}. \end{aligned} \tag{A}$$

Next, consider the false discovery rate for the HARKing protocol M^h. Recall that, in HARKing, a hypothesis is selected for reporting at random after observing the data from the set of hypotheses that are significant (if the set is nonempty).

Let π, α and β be as before, let p_ℓ denote the p-value of hypothesis with index ℓ, and let z_ℓ be a random variable such that $z_\ell = 1$ if hypothesis ℓ is selected to be reported and $z_\ell = 0$ otherwise. We obtain the false discovery rate for a hypothesis H^ℓ selected via HARKing as follows. Note that, by stipulation, if H^ℓ was selected for reporting, then it was in the set of significant hypotheses. We have

$$FDR(M^h) = \frac{Pr(p_\ell \leq \alpha, H_0^\ell, z_\ell = 1)}{Pr(p_\ell \leq \alpha, z_\ell = 1)}.$$

The likelihood-prior products can be expanded to $Pr(z_\ell = 1|p_\ell \leq \alpha, H_j^\ell)Pr(p_\ell \leq \alpha|H_j^\ell)Pr(H_0^\ell)$ for $j = 0, 1$. Note that the probability that H^ℓ is significant is independent of its truth conditional on its p-value. Hence, we have $Pr(z_\ell = 1|p_\ell \leq \alpha, H_j^\ell) = Pr(z_\ell = 1|p_\ell \leq \alpha)$, which cancels out in the numerator and (expanded) denominator. This leaves

$$\begin{aligned} &= \frac{Pr(p_\ell \leq \alpha|H_0^\ell)Pr(H_0^\ell)}{Pr(p_\ell \leq \alpha|H_0^\ell)Pr(H_0^\ell) + Pr(p_\ell \leq \alpha|H_1^\ell)Pr(H_1^\ell)} \\ &= \frac{\alpha(1-\pi)}{\alpha(1-\pi) + (1-\beta)\pi} \\ &= \Big(1 + \frac{\pi}{1-\pi}\frac{1-\beta}{\alpha}\Big)^{-1}. \end{aligned} \tag{B}$$

Now, compare equations (A) and (B), capturing the false discovery rates of prediction and HARKing protocols, respectively. When the fraction of true hypotheses selected for testing under prediction is the same as the fraction of true possible hypotheses, $\pi_p = \pi$, equations (A) and (B) are equal, and hence $FDR(M^p) = FDR(M^h)$, as desired. □

Proof of Proposition 2. To consider the false discovery rate of selective HARKing M^{sh}, let ℓ denote the the statistically significant hypothesis with the lowest p-value selected from the set of L hypotheses. That is, ℓ is the index of the hypothesis reported by selective HARKing. Let $-\ell'$ be the index of the hypothesis with the next-lowest p-value $p^*_{-\ell} = \inf_{\ell' \neq \ell} p_{\ell'}$.

The false discovery rate of selective HARKing then is just equal to the expectation of the false discovery rate of the hypothesis reported by selective HARKing, $FDR(M^{sh}) = \mathbb{E}_\ell[FDR(M_\ell^{sh})]$. Further, the false discovery rate of the selected hypothesis is equal to its expectation under the distribution of the p-values of the hypothesis with the next-lowest p-value $FDR(M_\ell^{sh}) = \mathbb{E}_{p^*_{-\ell}}[FDR(M_\ell^{sh}(p^*_{-\ell}))]$. Deriving the false discovery rate of $M_\ell^{sh}(p^*_{-\ell})$ just as before we get

$$\begin{aligned} FDR(M_\ell^{sh}(p^*_{-\ell})) &= \Big(1 + \frac{\pi}{1-\pi}\frac{1-\beta(\min\{\alpha, p^*_\ell\})}{\min\{\alpha, p^*_\ell\}}\Big)^{-1} \\ &\leq \Big(1 + \frac{\pi}{1-\pi}\frac{1-\beta}{\alpha}\Big)^{-1} = FDR(M^p) \end{aligned}$$

Thus $FDR(M^{sh}) \leq FDR(M^h) = FDR(M^p)$ as desired.

To prove the false discovery rate for selective HARKing is decreasing in the number of hypotheses, first note that $P_\ell^{sh}(p^*_{-\ell})$ is increasing in $p^*_{-\ell}$ for $p^*_{-\ell} < \alpha$ and constant otherwise. The CDF of $p^*_{-\ell}$ given the number of hypotheses L is $Pr(p^*_{-\ell} \leq t|L) = Pr(p^*_{-\ell} \leq t)^{L-1}$ and thus $Pr(p^*_{-\ell} \leq t|L)$ is first-order stochastically dominated by $Pr(p^*_{-\ell} \leq t|L')$ for $L' < L$. It follows that P^{sh} is decreasing in L. □

Proof of Proposition 3. Finally, we show that prediction is more reliable than HARKing, $FDR(P^p) > FDR(P^h)$ just in case $\pi_p > \pi$. Consider a population of studies and let π be the fraction of true hypotheses in the set of possible hypotheses; let π_p be the fraction of true hypotheses in the subset of hypotheses selected via prediction; α the significance threshold of tests; and $1-\beta$ their power.

From the preceding proofs, we have the following false discovery rates for the prediction and HARKing protocols:

$$FDR(M^p) = \Big(1 + \frac{\pi_p}{1-\pi_p}\frac{1-\beta}{\alpha}\Big)^{-1}, \text{ and } FDR(M^h) = \Big(1 + \frac{\pi}{1-\pi}\frac{1-\beta}{\alpha}\Big)^{-1}.$$

So $FDR(M^p)$ and $FDR(M^h)$, as established in proposition 2, are equal when the fraction of true hypotheses selected under either protocol are equal, $\pi_p = \pi$. Thus, all that remains to be shown is that the false discovery rate of prediction is decreasing in π_p. For this, we simply take the derivative of $FDR(M^p)$ with respect to π_p and show that it is negative.

$$\frac{\partial}{\partial \pi_p}[FDR(M^p)] = \frac{\alpha(\beta-1)}{\big(\alpha(\pi_p-1)+(\beta-1)\pi_p\big)^2} < 0.$$

And the expression is negative since the numerator is negative for the assumed values of type I and II error rates ($\alpha, \beta \in (0,1)$) and since the denominator must be positive. □

Proposition 4. *When hypotheses are selected from the same set of candidate hypotheses with fraction $\pi \in (0,1)$ true hypotheses, then prediction yields the same false omission rate as HARKing.*

Proof. Let π_p, π, p, α, and β be as before. Consider the false omission rate of prediction, M^p.

$$\begin{aligned} FOR(M^p) &= \frac{Pr(p > \alpha, H_1)}{Pr(p > \alpha)} \\ &= \frac{Pr(p > \alpha | H_1)Pr(H_1)}{Pr(p > \alpha | H_1)Pr(H_1) + Pr(p > \alpha | H_0)Pr(H_0)} \\ &= \frac{\beta\pi_p}{\beta\pi_p + (1-\alpha)(1-\pi_p)} \\ &= \Big(1 + \frac{1-\pi_p}{\pi_p}\frac{1-\alpha}{\beta}\Big)^{-1}. \end{aligned} \tag{A}$$

Next, consider the false omission rate of the HARKing protocol, M^h.

$$\begin{aligned} FDR(M^h) &= \frac{Pr(p_\ell \leq \alpha, H_0^\ell, z_\ell = 1)}{Pr(p_\ell \leq \alpha, z_\ell = 1)}. \\ &= \frac{Pr(p_\ell \leq \alpha | H_0^\ell)Pr(H_0^\ell)}{Pr(p_\ell \leq \alpha | H_0^\ell)Pr(H_0^\ell) + Pr(p_\ell \leq \alpha | H_1^\ell)Pr(H_1^\ell)} \end{aligned}$$

$$= \frac{\alpha(1-\pi)}{\alpha(1-\pi)+(1-\beta)\pi}$$
$$= \Big(1+\frac{1-\pi}{\pi}\frac{1-\alpha}{\beta}\Big)^{-1}. \tag{B}$$

Clearly, whenever $\pi_p = \pi$, the false omission rates given in (A) and (B) are equal, as desired. □

Proposition 5. *Prediction is more reliable than HARKing with respect to false omission, $FOR(P^p) > FOR(P^h)$, just in case $\pi_p < \pi$.*

Proof. Let π_p, π, p, α, and β be as before. Consider the false omission rate of prediction, M^p. From the preceding proofs, we have the following false omission rates for the prediction and HARKing protocols:

$$FOR(M^p) = \Big(1+\frac{1-\pi_p}{\pi_p}\frac{1-\alpha}{\beta}\Big)^{-1}, \text{ and } FOR(M^h) = \Big(1+\frac{1-\pi}{\pi}\frac{1-\alpha}{\beta}\Big)^{-1}.$$

Now, observe that false omission rate of prediction is increasing in π_p. For this, take the derivative of $F0R(M^p)$ with respect to π_p and show that it is positive.

$$\frac{\partial}{\partial\pi_p}[FOR(M^p)] = \frac{\beta(1-\alpha)}{\left(\alpha(\pi_p-1)+(\beta-1)\pi_p+1\right)^2} > 0,$$

Since the numerator is positive (given $\alpha > 0$) as is the denominator. □

Computational Appendix

A GUI for exploring the performance of reporting protcols for prediction, HARKing, and fallback HARKing is available at: `https://amohseni.shinyapps.io/Reporting-Protocols-and-the-Reliability-of-Science/`.

All the R code for this project is available at GitHub at: `https://github.com/amohseni/Reporting-Protocols-and-the-Reliability-of-Science`.

References

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka,

K. Imai, G. Imbens, J. P. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson (2018). Redefine statistical significance. *Nature Human Behaviour*.

Benjamini, Y. and Y. Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*.

Bird, A. (2018). Understanding the Replication Crisis as a Base Rate Fallacy. *The British Journal for the Philosophy of Science*.

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*.

Bright, L. K. (2017). On fraud. *Philosophical Studies*.

Bruner, J. P. and B. Holman (2019). Self-correction in science: Meta-analysis, bias and social structure. *Studies in history and philosophy of science*.

Dreber, A., T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, M. Johannesson, and K. W. Wachter (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences of the United States of America*.

Etz, A. and J. Vandekerckhove (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*.

Franklin, A. (2013). *Shifting standards: Experiments in particle physics in the twentieth century*. University of Pittsburgh Press.

Good, I. J. (1967). On the principle of total evidence. *British Journal for the Philosophy of Science*.

Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions (2015). The Extent and Consequences of P-Hacking in Science. *PLoS Biology*.

Heesen, R. (2018). Why the reward structure of science makes reproducibility problems inevitable.

Hitchcock, C. and E. Sober (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*.

Hollenbeck, J. R. and P. M. Wright (2016). Harking, Sharking, and Tharking: Making the Case for Post Hoc Analysis of Scientific Data. *Journal of Management*.

Ioannidis, J. P. (2018). The proposal to lower P value thresholds to .005.

John, L. K., G. Loewenstein, and D. Prelec (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*.

Lehmann, E. L. (1993). The fisher, Neyman–Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*.

Lyons, L. (2013). Discovering the Significance of 5 Sigma. *arXiv*.

Lyons, L. (2015). Statistical Issues in Searches for New Physics. *arXiv*.

Machery, E. (2020). What is a replication? *Philosophy of Science*.

Mayo, D. (2019). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press.

Mayo-Wilson, C. and S. C. Fletcher (2019). Evidence in Classical Statistics. In M. Lasonen-Aarnio and C. Littlejohn (Eds.), *Routledge Handbook of the Philosophy of Evidence*. Routledge.

Mill, J. S. (1843). *A System of Logic, Ratiocinative and Inductive*. Routlege.

Murphy, K. R. and H. Aguinis (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology*.

Pawel, S. and L. Held (2020). Probabilistic forecasting of replication studies. *PLoS ONE*.

Popper, K. (1934). *The logic of scientific discovery*. Routlege.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass*.

Romero, F. (2020). The Division of Replication Labor. *Philosophy of Science*.

Romero, F. and J. Sprenger (2020). Scientific self-correction: the Bayesian way. *Synthese*.

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*.

Savage, L. (1954). *The Foundations of Statistics*. Wiley Publications in Statistics.

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

Szucs, D. and J. P. Ioannidis (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*.

Tetlock, P. E. (2017). *Expert political judgment: How good is it? How can we know?, new edition*. Princeton University Press.

Wassertheil, S. and J. Cohen (2006). Statistical Power Analysis for the Behavioral Sciences. *Biometrics*.

University of California, Irvine; Irvine, CA 92697, USA
Email address: amohseni@uci.edu