Preregistration and Predictivism

Hong Hui Choi

University of Pittsburgh, Department of History and Philosophy of Science, PA, USA

Forthcoming at *Synthese*.

1. Introduction

In recent years, several scientific disciplines have been undergoing replication crises, and in response, preregistration has been offered as a solution to replicability problems. In this paper, I will draw connections between this new focus on preregistration and an older debate in the philosophy of science, namely predictivism—the thesis that predictions are epistemically superior to accommodations. Specifically, I shall argue that predictivism justifies preregistration. As it turns out, predictivists of all stripes have subtly different reasons to support preregistration. This unity is significant because firstly, preregistration proponents often seem to be implicitly committed to stronger versions of predictivism, but strong predictivism has long been deemed untenable by philosophers of science. Furthermore, the efficacy of preregistration in dealing with Questionable Research Practices (QRPs) like p-hacking and HARKing is at best contentious, which blocks a straightforward empirical justification of preregistration. Although empirical validation of preregistration—which need not be based on preventing QRPs—will eventually be required, it would be nice to have a principled justification for preregistration while we await the empirical evidence. I will also argue that preregistration offers something in return to predictivism: the former bolsters the latter by serving as a counterargument to accommodationism—the antithesis of predictivism.

The plan is as follows: In section 2, I provide the rational for preregistration by briefly explaining the history of the replication crisis; in section 3, I make the predictivist argument for preregistration by showing that different versions of predictivism each has subtly different reasons to support preregistration; in section 4, I show that preregistration returns the favor to predictivism by suggesting that *preregistered* predictions serves as a counterargument to Finnur Dellsén's accommodationism; finally, in section 5 I conclude by reiterating the main points.

2. Replication Crisis

To explain the rationale behind preregistration, I will first give a quick history of the so-called replication crisis.[1] In recent years, researchers started worrying that many of their studies are unreliable and cannot be replicated. There are many reasons why such worries arose; I will mention just two.[2] Firstly, Daryl Bem's paper *Feeling the Future* was published in the prestigious *Journal of Personality and Social Psychology*, where he allegedly found evidence for extrasensory perception (Bem 2011). Bem's publication suggested to many that there were serious problems with researchers' methodologies and journals' publication strategies.

[1] The examples cited in this paper are mostly drawn from psychology, but the connections between preregistration and predictivism I am arguing for should generalize beyond psychology.

[2] In fact, such worries are not even new, at least in psychology: Lakens (2023) has argued that the contemporary crisis is similar to another crisis from the 1960s.

Secondly, the worry over problematic methodology was supported by surveys such as John et al. (2012) which found high prevalence of shoddy statistical practices—commonly called Questionable Research Practices (QRPs) in psychology. Notable examples of QRPs include p-hacking which are bad statistical practices aimed at achieving statistical significance (Simonsohn et al. 2014) and Hypothesising After Results Known (HARKing) which are accommodations disguised as predictions (Kerr 1998).[3]

In response to such worries, several disciplines undertook massive replication projects to assess the replicability of their field. For example, in psychology, the Open Science Collaboration (OSC) attempted to replicate 100 papers published in highly ranked psychology journals, (OSC 2015). Other disciplines that undertook replication projects include economics (Camerer et al. 2016), cancer biology (Errington et al. 2021), and even the nascent subfield of experimental philosophy (Cova et al. 2021). Many of these replication projects estimated their discipline's replicability to be poor: less than 50% for psychology (OSC 2015) and cancer biology (Errington et al. 2021). Camerer et al. (2016) estimated 61% for economics but the replication effect sizes were on average only 66% that of the original (Camerer et al. 2016). An exception is experimental philosophy, which estimated replicability to be around 70% and that replication effect sizes were not smaller than the originals. Nevertheless, the general sentiment is that many disciplines are grappling with replication problems, leading to a "crisis of confidence" in science (Earp and Trafimow 2015).

Thereafter, preregistration—which refers to specifying one's research plans in a repository prior to data collection and analysis—was touted as a revolutionary solution to the replication crisis (Nosek et al. 2018). According to this narrative, preregistration is supposed to prevent QRPs by acting as an enforcement tool to distinguish between predictions and accommodations disguised as predictions (*ibid*, pp.2601-2602). Registered Reports is a specific form of preregistration where researchers submit their preregistrations for peer review and receive in-principle-acceptance before data is collected. After data collection, researchers submit their full manuscripts (including the raw data set) for peer review again to ensure compliance with their preregistrations. In addition to preventing QRPs, Registered Reports are supposed to have the added benefit of remedying publication bias as journals are committed to publish the research regardless of how the results turn out (Chambers and Tzavella 2022).

Critics of preregistration have tended to focus on whether there is empirical evidence of its efficacy.[4] Indeed, there is some evidence against the efficacy of preregistrations in preventing QRPs and publication bias. To effectively constrain QRPs, there must be minimal deviation between the preregistration and the published (or submitted) manuscript.[5] Unfortunately, there is evidence that deviations are common in preregistered studies (Heirene et al. 2024; Claesen et al. 2021). In particular, van den Akker et al. (2023, p.30) found that sampling plans and statistical analyses are most likely to deviate compared to other parts

[3] "QRPs" is a vaguely defined term meant to cover a wide range of 'bad' practices. Some even include fraud as an example of a QRP (John et al. 2012, p.525). In this paper, I focus on p-hacking and HARKing as they are most relevant to preregistration.

[4] Some critics have also suggested that preregistration distracts from the bigger problem of underdeveloped theories (Szollosi et al. 2020; van Rooij and Baggio 2021; and Szollosi and Donkin 2021).

[5] More on this in section 3.1.

of a study. Furthermore, there is also evidence that preregistered studies are not immune to publication bias. Ensinck and Lakens (2023) estimated that close to half of the preregistrations on the Open Science Framework's (OSF) repository remain unpublished. [6] Worryingly, Ensinck and Lakens estimated that a quarter of the preregistrations remain unpublished because the authors failed to find clear positive results (ibid, p.5); this suggests that insufficient lesson has been drawn from the decades of discussion about the harms of publication bias.

On the other hand, preregistration proponents have responded with empirical evidence *for* the efficacy of preregistrations. Protzko et al. (2023) claimed to show in a prospective study that rigour-enhancing methods—including preregistration—contributed to the discovery of highly replicable findings. However, Protzko et al. (2023) have been criticized for overstating their claims. Most notably, Bak-Coleman and Devezer argued that Protzko et al. (2023) used unusual definitions of replicability and did not employ control conditions which then precludes them from making causal claims (Bak-Coleman and Devezer 2024, p.1891). Furthermore, in an ironic twist, it appears that Protzko et al. (2023) did not themselves preregister their main causal claim regarding preregistration's efficacy (Bak-Coleman and Devezer 2024; p.1890).[7] Lastly, another study by van den Akker and colleagues failed to find evidence that preregistration prevents p-hacking and HARKing (van den Akker 2023).

Nevertheless, there is more convincing evidence that Registered Reports are effective at constraining research flexibility and publication bias. Wiseman et al. (2019) and Scheel et al. (2021) each found that Registered Reports have a lower rate of positive results compared to non-preregistered studies. While there are many possible explanations for this disparity—such as the proportion of true hypotheses being tested, statistical power, and of course, unchecked flexibility and publication bias—Scheel et al. (2021, p.9) noted that if we assume that Registered Reports are totally inefficacious in remedying research flexibility and publication bias, then the disparity in positive rate would require that non-preregistered studies test >90% true hypotheses and have >90% statistical power—but both of these assumptions are implausible and contradicted by extant research (e.g., Szucs and Ioannidis 2017).

The upshot is that it is yet unclear whether there is good empirical evidence for preregistration's efficacy at ameliorating QRPs and publication bias. Though the efficacy of Registered Reports seems more promising, the controversy over empirical efficacy seems unlikely to be resolved anytime soon. Luckily for proponents of preregistration, an empirical justification can be set aside for a moment if we possess a principled defense of preregistration on firm theoretical grounds. I will argue now that predictivism provides such grounds.

3. The Predictivist Argument for Preregistration

Predictivism is the thesis that predictions are epistemically superior to accommodations. Before diving into the various versions of predictivism, we should first clarify the distinction

[6] It should be noted that Registered Reports were excluded in this analysis because they were deemed unlikely to be affected by publication bias (Ensinck and Lakens 2023, p.3).

[7] The concerns raised in Bak-Coleman and Devezer (2024) eventually led to the retraction of Protzko et al. (2023). See Protzko et al. (2024) for more details on the retraction.

between prediction and accommodation. As a first pass, the most natural reading seems to be a temporal one: a theory predicts the data only if the theory precedes the data; a theory accommodates the data only if the data precedes the theory. Despite its intuitiveness, the temporal distinction is widely rejected for two reasons. Firstly, it seems peculiar that the temporal relation between theory and data should have consequence on the data's confirmation of the theory. Secondly, the temporal distinction conflicts with actual scientific practice, as there are instances of data preceding the theory and yet the theory is widely recognized to be well-confirmed by the data. For example, Zahar (1973, p.101) argued against the temporal distinction by noting that it leads to a "paradoxical situation". The situation is that the anomalous precession of Mercury's perihelion temporally preceded Einstein's discovery of his theory of general relativity, and thus the theory accommodated that data according to the temporal distinction. Yet, it was widely recognized by the scientific community that Einstein's theory of general relativity was strongly confirmed by the anomalous precession of Mercury's perihelion.

In lieu of the temporal distinction, the use-novelty distinction enjoys wider adoption. The use-novelty distinction distinguishes prediction and accommodation according to whether the data was *used* to construct the theory: a theory predicts the data only if the data was not used to construct the theory; a theory accommodates the data only if the data was used to construct the theory. An advantage of the use-novelty distinction is that it accords better with the scientific record. Returning to the Einstein example, we see that the use-novelty distinction would say that Einstein's theory of general relativity predicted the anomalous precession of Mercury's perihelion because the data was not used to construct the theory. Thus, there is nothing strange about the scientific community's consensus that Einstein's theory was strongly confirmed.

A problem for the use-novelty distinction was proposed by Musgrave (1974, p.13).[8] Suppose that scientist A proposed theory T by accommodating data sets P and Q. Scientist B also independently proposed T by accommodating P only.[9] The troubling question is: does Q confirm T? It seems like it does and that it does not. Q does not confirm T since A's T accommodated Q, but Q confirms T since B's T predicted Q. We can rescue ourselves from contradiction by claiming that Q does not confirm T *for A* but Q confirms T *for B*. But this seems highly unintuitive as we might think that confirmation is more objective than this. Leplin (1997, p.54) pressed a similar worry by noting that "[t]he theorist's hopes, expectations, knowledge, intentions, or whatever, do not seem to relate to the epistemic standing of his theory in a way that can sustain a pivotal role for them." The problem that Musgrave and Leplin are pressing is essentially that the use-novelty distinction suggests that confirmation of theory turns on the mental states of the investigator. The driving force of this intuition is perhaps an implicit belief that confirmation is a purely logical relation between theory and evidence. However, Musgrave's and Leplin's arguments are not decisive blows against the use-novelty distinction. John Worrall (1985, 2005, 2006) has developed a more sophisticated version of a use-novelty account that moves away from *actual* usage of data and towards

[8] Musgrave (1974) had his own take on the prediction-accommodation distinction. I will not consider it here because firstly, Musgrave's distinction does not enjoy much uptake in the predictivism debate and secondly, it faces problems of its own (Worrall 2006, pp.36-38).

[9] Let us set aside the implausibility of such a scenario taking place. As far as I know, no pair (or more) of scientists have found themselves in such a curious situation.

whether data was *needed* to construct the theory. As we shall see in section 3.5, this move allows Worrall to avoid Leplin's and Musgrave's concerns.

Furthermore, it was also Musgrave (1974, pp.3-8) who noted that pure logical confirmation faces paradoxes of its own. Lastly, unlike philosophers, psychologists do not seem to find the notion that confirmation turns on psychological states mysterious. A common distinction made by psychologists is between Confirmatory Data Analysis (CDA) and Exploratory Data Analysis (EDA). CDA and EDA are sets of tools psychologists use in their research and they are defined in terms of researcher *intention*: psychologists engage in CDA when they use data to *test* theories or hypotheses while they engage in EDA when they use data to *discover* hypotheses (Fife and Rodgers 2022, p.462). A cardinal rule of EDA is that a set of data used under EDA to discover hypotheses may not be used again for CDA to test the discovered hypotheses. Another rule of EDA is that its "findings are provisional, awaiting confirmation [from] an independent dataset" (*ibid*, p.458). Psychological practice thus appears to vindicate the 'mysterious' use-novelty distinction. In any case, my primary purpose in this paper is not to argue for a particular prediction-accommodation distinction, so I will assume the use-novelty distinction just because it is most widely used in the predictivism debate. I will now argue that different versions of predictivism provide subtly different justifications for preregistration.

3.1 Confirmatory Strong Predictivism

Strong predictivism says that the epistemic superiority of predictions is an *inherent* quality of prediction. One influential version of strong predictivism is from Ronald Giere, who argued that accommodations provide zero confirmation:

> If the known facts were used in constructing the model and were thus built into the resulting hypothesis…then the fit between these facts and the hypothesis provides no evidence that the hypothesis is true [since] these facts had no chance of refuting the hypothesis (Giere 1984, p.161).

Although Giere's strong predictivism is probably a minority view among philosophers of science, it is worth noting that some proponents of preregistration seem to be committed—if only implicitly—to some version of strong predictivism. For example, Nosek and colleagues seem to hold that predictions and accommodations are epistemically unequal:

> Why does the distinction between prediction and postdiction matter? Failing to appreciate the difference can lead to overconfidence in post hoc explanations (postdictions) and inflate the likelihood of believing that there is evidence for a finding when there is not (Nosek et al., p.2600).

While Nosek and colleagues did not explicitly commit themselves to strong predictivism, some statements they made are rather suggestive. For example, their quote from above suggests that they think that predictions provide evidence for a finding while accommodations do not. This is not too far from Ronald Giere's version of strong predictivism, which holds that predictions are confirmatory while accommodations are not (Giere 1984).

One may wonder why preregistration is needed to distinguish prediction from accommodation. The reason is due to the prevalence of QRPs like p-hacking and HARKing, as mentioned above. Notice that what unites QRPs like p-hacking and HARKing is that they are accommodations *disguised* as predictions. A p-hacker uses a range of shoddy methods in order to achieve statistical significance. For example, he might employ multiple independent variables or measure multiple dependent variables. He might also split his data set into various subsets and conduct statistical testing on all of them. By combining these strategies, a p-hacker is virtually guaranteed to achieve statistical significance. However, after waddling through the garden of forking paths and finally emerging in the promised land of $p < .05$, p-hackers will not transparently report the numerous twists and turns they traversed through. Instead, they report a straight path from hypothesis to statistical significance. Of course, the p-hacker's 'prediction' was data dependent, thus p-hacking is accommodation disguised as prediction. HARKing is even more obviously a case of accommodations disguised as predictions: a HARKer constructs a hypothesis after observing the data but claims to have predicted the data. Preregistration enables one to distinguish genuine predictions from disguised accommodations as the latter will inevitably result in deviations between the commitments made in the preregistration and the eventual manuscript. In the context of prevalent p-hacking and HARKing, strong predictivists like Giere and Nosek and colleagues should welcome preregistration as it serves as a practical enforcement tool to distinguish predictions from accommodations disguised as predictions. This distinction is important for strong predictivists since they hold that there is asymmetry between predictions and accommodations with regards to confirmability.

Of course, preregistration's ability to enforce a practical distinction between prediction and accommodation applies to all versions of predictivism. However, I will show next that different versions of weak predictivisms—which says only that prediction is *indicative* of other epistemic qualities—all have subtly different reasons to support preregistration. To anticipate, the general reason is that preregistration is informative about the different epistemic qualities thought to be important according to the various weak predictivisms.

3.2 Maher's Weak Predictivism

Patrick Maher's (1988) weak predictivism identified the *reliability of the method that produced the hypothesis or theory* as the epistemic quality that prediction is indicative of. Maher's argument provided a Bayesian justification for why prediction is epistemically superior to accommodation. The idea is that a theory that successfully predicted the data is indicative that the theory was produced by a (more) reliable method compared to a theory that simply accommodated the same data. In turn, knowledge of the method's reliability leads to a higher posterior belief in the theory's truth, compared to when such knowledge is absent. Maher (1988, pp.280-281) provided a rigorous proof for why successful prediction implies reliable method which in turn implies higher posterior belief. For my purposes, the following thought experiment, also proposed by Maher, suffices to illustrate the argument:

> A coin is flipped 99 times and the sequence of heads and tails is seemingly random. Peter did not observe the flips but was able to successfully predict all 99 flips. Adel observed the flips and reported the results of all 99 flips. Subsequently, both Peter and Adel independently predict that the 100^{th} flip will land heads. The conjunction of

all 100 flips (the theory) is logically equivalent between Peter and Adel. What should our posterior belief in Peter's and Adel's theories be?

The intuition that Maher was suggesting is that our posterior belief in Adel's theory should be 0.5 while our posterior belief in Peter's theory should be close to 1 (*ibid*, p.275). Since Adel's and Peter's theories are logically equivalent, what explains the difference in our posterior beliefs? Maher's suggestion was that Peter's 99 successful predictions indicate that his theory was produced by a reliable method while Adel's 99 accommodations say nothing about the reliability of his method. That is, the knowledge that Peter possesses a reliable method explains our higher posterior belief in his theory compared to Adel's theory.

The question then, for my purposes, is whether preregistration provides information about methodological reliability. Maher defined reliability of the method as the objective probability that a theory is true, given that it was generated by said method (p.276). In other words, methodological reliability is the empirical track record of theories proposed by the method: the proportion of true theories among all theories proposed by that method. Notice that "method" need not indicate any concrete algorithm that 'discovers' theories. Rather, we may speak loosely of a given researcher's 'method' to generate theories, which may be anything from literature review to dream inspiration. The unit of analysis also need not be individual researchers, but also research labs, universities, disciplines, or any configuration of researchers. Regardless of whose method, what is clear is that our estimate of reliability can be highly misleading if we do not possess a *representative* sample of the method's track record. For example, if we only have access to true theories proposed by a method but not false theories, we may be misled into thinking that the method is reliable even if it proposed many more false theories than true theories.

The scientific record is known to be distorted by biases like the file-drawer problem (where negative results are not submitted to journals) (Rosenthal 1979) and publication bias (where positive results are more likely to be published by journals than negative results) (Franco et al. 2014). Preregistration provides protection against such distortions of a method's track record. By evaluating all the preregistrations posted by a unit, we have a more representative sample of the method's reliability compared to relying only on published papers. Of course, the preregistration record is not guaranteed to be undistorted, since there are ways to circumvent it, such as preregistering after results known (PARKing) (Yamada 2018). However, the point is not that preregistration is fool proof but rather that it provides some protection against distorted assessments of a method's reliability. This latter, weaker requirement is all we need to secure knowledge that a method is reliable, which in turn secures higher posterior belief in its theories.

3.3 Lange's Weak Predictivism

Marc Lange's (2001) weak predictivism identified the *unarbitrariness of conjunctions* as the epistemic quality that prediction is indicative of. Lange's weak predictivism drew heavily on Maher's coin flip example; the difference is that Lange argued that Maher had misdiagnosed the epistemic quality that explains why successful predictions correlate with superior confirmation. Recall that Maher's example involves a seemingly random sequence of 99 coin flips, and the moral there was that Peter's theory (that predicted the flips) is well-confirmed while Adel's theory (that accommodated the flips) is not well-confirmed because the former's

successful predictions indicate a reliable method. Lange agrees that Maher's account gets it right in this example, but gets it wrong in a slightly tweaked example where the 99 coin flips form a strictly alternative (heads, tails, heads, tails,…) rather than a random sequence (Lange 2001, p.580). In this tweaked example, Lange argued that intuitively both Peter's and Adel's theories are equally well confirmed. Importantly, in this tweaked example, it is inconsequential whether the flips were predicted or accommodated or whether the method is reliable or unreliable—thus Maher's account gets it wrong (*ibid*, p.582).

Rather than reliability of method, Lange proposed that the arbitrariness of the conjunction (of coin flips) explains our intuitions (p. 581). In Maher's example involving a random sequence of coin flips, the conjunction of the 99 coin flips appears arbitrary while in Lange's tweaked example involving a strictly alternating sequence, the conjunction is not arbitrary. Furthermore, accommodation of arbitrary conjunctions does not usually lead to strong confirmation since the conjuncts "may well lose any connection to one another" (p.583). The arbitrariness of the conjunction of coin flips explains why accommodation is not well confirmed in Maher's example (involving an arbitrary conjunction) but accommodation is well confirmed in Lange's example (involving an unarbitrary conjunction). The final piece in Lange's account is that predictions correlate with unarbitrary conjunctions while accommodations correlate with arbitrary conjunctions, thus the former is indicative of epistemic superiority over the latter due to its correlation with the arbitrariness of conjunction (pp.583-584).

The question then, for my purposes, is whether preregistration provides information about the arbitrariness of conjunctions. Lange did not offer a precise definition of arbitrariness but his idea that arbitrary conjunctions are characterized by unrelated conjuncts seem on the right track. Lange cited Goodman's example of an arbitrary conjunction as "8497 is a prime number and the other side of the moon is flat and Elizabeth the First was crowned on a Tuesday" (Goodman 1983, pp.68-69). For an actual example from science, consider Daryl Bem's (2011) infamous studies that purported to find evidence of extrasensory perception (ESP)—roughly, the idea that humans have a sixth sense that defies known physical laws due to its retroactive nature. Bem (2011) discussed 9 experiments, all of which involved testing for ESP; but a curiosity is the heterogeneous types of tests employed. For example, experiment 1 involved erotic stimuli; experiment 2 involved negative stimuli; experiment 3 involved the priming effect; experiment 5 involved the habituation effect; etc. One wonders why *these* seemingly unrelated tests were reported; that is, why did Bem present this seemingly *arbitrary* conjunction of experiments as evidence? One compellingly simple answer, of course, is that these experiments were selected based on statistical significance rather than a transparent reporting of the *entire* conjunction of experiments. That is, the results were p-hacked.

Bem (2011) is an example where we can detect an arbitrary conjunction even without access to the entire conjunction, but this may not always be so easy—especially if our priors were not already automatically reduced by the ESP subject matter. In subtler cases, preregistration can help detect arbitrary conjunctions as it allows us to compare the entire conjunction of planned observations found in the preregistration against the reported conjunction of observations found in the submitted manuscript. To borrow Lange's example, suppose that a manuscript reported the following conjunction of coin flips (in temporal order):

> Heads, Tails, Heads, Tails, Heads, Tails, Heads, Tails, Heads, Tails, Heads, Tails, Heads, Tails, Heads, Tails, Heads, Tails, Heads, Tails.

According to Lange's account, this conjunction of flips is not arbitrary and thus strongly confirms the relevant theory. However, suppose that the preregistration of this manuscript had planned for 35 coin flips but only the 20 above were eventually reported. This should alert us to the possibility that the seemingly unarbitrary conjunction that was reported was selected from the actual, arbitrary conjunction of observations:

> **Actual, arbitrary conjunction**: HHTHHTTHTHHHTHHTTHHHHHTHTHHHTHTHHTT
> **Omitted flips underlined**: HHTHHTTHTHHHTHHTTHHHHHTHTHHHTHTHHTT
> **Reported, 'unarbitrary' conjunction**: HTHTHTHTHTHTHTHTHTHT

3.4 Mayo's Weak Predictivism

Deborah Mayo's theory of severity has a negative and a positive aspect. According to the negative aspect, there is little to no evidence for a theory if it passes a test that has little to no capability of finding flaws with the theory (Mayo 2018, p.5). According to the positive aspect, there is strong evidence for a theory if it passes a test that is highly capable of finding flaws with the theory (if it is flawed). Mayo has argued that the predictivist intuition that predictions are superior to accommodations is a result of it correlating with severity (Mayo 1991). That is, Mayo's weak predictivism identified *test severity* as the epistemic quality that prediction is indicative of.

What is the relevance of preregistration to Mayo's weak predictivism? Daniel Lakens has argued that the function of preregistration is to allow others to transparently evaluate the severity of a test (Lakens 2019; Lakens et al. 2024). The idea here is that the information contained in a preregistration is informative of the severity of the test that the theory underwent. Importantly, severity can be difficult to assess if the researcher did not make prior commitments before the test (i.e., preregistered their study). For example, in HARKing, researchers construct a theory after analyzing the data. Of course, barring unintended errors, a theory constructed through HARKing is guaranteed to fit well with the data; that is, the test will not be severe. Obviously, HARKers will not openly admit that their theory was constructed using the data; instead, they will claim that the theory produced successful predictions. Disguised accommodations and genuine predictions will be difficult to tell apart unless the researcher was forced to specify the theory and its associated predictions without access to the data—precisely what a preregistration does.

Lakens' argument is on the right track, but it is incomplete. His argument tells us that the *sociological* function of preregistration is to allow researchers to transparently evaluate the severity of a manuscript, but we may wonder why *that* is epistemically significant. To make the sociological-epistemic distinction clear, consider peer review as an example. A sociological function of peer review may be that there must be some sort of manuscript selection process since the available space for publication is almost always smaller than the number of submitted manuscripts. Or perhaps a sociological function of peer review is that it lends credibility to published manuscripts. In contrast, an epistemic function of peer review may be that the vetting of manuscripts provided by experts prevents low quality submissions

from entering the published record and possibly misleading others. One suggestion for Lakens is to borrow Helen Longino's theory of science as social knowledge (Longino 1990, 2002). Specifically, Longino has argued that scientific observation is fundamentally social in nature:

> The claim of sociality is the claim that the status of the scientist's perceptual activity as observation depends on her relations with others, in particular *her openness to their challenge to and correction of her reports* (Longino 2002, p.103, added emphasis).

In other words, what makes a piece of data a *scientific observation* as opposed to individual sensory perceptions is an open and transparent attitude to be corrected (or supported) by others. In a similar vein, the sociological function of preregistration—to allow others to transparently evaluate severity—can be transformed to become an epistemic function since the very same transparency to evaluation *is* what makes the content in a preregistration scientific.

3.5 Worrall's Predictivism

John Worrall has propounded predictivism for decades and perhaps because of this, his views have undergone subtle changes. Samuel Schindler helpfully distinguished weak and strong versions of Worrall's predictivism (Schindler 2014, p.63; Schindler 2018, p.73)[10]. In the weak version, Worrall's understanding of use-novelty follows standard usage, which is based on whether the data was *actually used* to construct the theory:

> because of the way [the theory that accommodated the data] was constructed, it was never at risk of refutation from these observations. *This fact cannot be discovered by simply inspecting the logical form of the theory* (Worrall 1985, p.313, original emphasis).

But confusingly, Worrall seemed to also have in mind a stronger version which concerns whether the data was *needed* to construct the theory:

> [I]t is no part of the heuristic view that it should matter what Einstein was worrying about at' the time he produced his theory, what matters is only whether he *needed* to use some result about Mercury in order to tie down some part of his theory (*ibid*, p.319, added emphasis).

An advantage that the strong version has over the weak version is that it helps Worrall avoid Musgrave's (1974) and Leplin's (1997) criticisms that confirmation should *not* turn on the psychological states of researchers. Hence, in later writing, Worrall took pains to insist on the strong version:

[10] It should be noted that the weak and strong versions of *Worrall's* predictivism are distinct from the more *general* distinction between weak and strong predictivism.

> My account gives no role to any such psychological factor. Although presented as a version of the 'heuristic approach', it is at root a *logical* theory of confirmation (Worrall 2005, p.819, original emphasis).[11]

Keeping in mind the foregoing will be key to understanding Worrall's predictivism, where he distinguished between *conditional* confirmation and *unconditional* confirmation. Worrall's framework involves a general theory T, a piece of data D, and a specific theory T*. In conditional confirmation a researcher in possession of T uses D to construct T*. However, Worrall rightly noted that conditional confirmation can be trivially achieved. For example, when T* just is the conjunction of T and D. In such a case, T* entails D but it does not seem like D non-trivially confirms T* since this strategy allows D to 'tack on' to *any* arbitrary T (Worrall 1985, p.302). The tacking paradox teaches us that the *choice* of T* matters in confirmation. Yet, even when there is an unarbitrary connection between T* and D, the confirmation that result can nevertheless be *conditional* on prior acceptance of T. Worrall's favorite example to illustrate this is the "Gosse dodge", referring to Phillip Gosse's *Omphalos* theory that God created fossil records that *look* millions of years old despite the Creationist thesis that the Earth is only thousands of years old. Worrall's point here is that Goose's specific Omphalos theory (T*) is confirmed by the fossil record (D) only if the general Creationist theory (T) is *already accepted*. The fossil records are evidence that Creationists should believe Gosse's Omphalos theory, but they are *not* evidence for non-Creationists such as Darwinian Evolutionists to believe in Creationism. That is, D confirms T* *conditional* on prior acceptance of T and importantly, D does *not* confirm T (Worrall, 2005, p.817-818; Worrall 2006, p.44-45).

In contrast, in unconditional confirmation, the general theory T *does* receive confirmation. Worrall mentioned two categories of unconditional confirmation (Worrall 2005, p.818; Worrall 2006, p.47-51). Firstly, we saw that when a researcher in possession of T uses D to construct T*, then there is no confirmation of T; but when T* makes an *independent* prediction D* that turns out successful, then T and T* are both confirmed. Isaac Newton's theory of gravitation during the 19th century is a classic example of this. The astronomer John Adams had proposed that the (then) anomalous motion of Uranus could be resolved as such:

> According to my calculations, the observed irregularities in the motion of Uranus may be accounted for by supposing the existence of an exterior planet (Adams 1896, p.1).

What Adams did was to apply a correction to the calculations for the motion of Uranus using gravitational theory—the correction being the addition of the posited planet—and he showed that Newton's theory accommodated Uranus' motion relatively well if we included the gravitational impact of the posited planet (*ibid*, p.4-5). In other words, Adams, in possession of Newton's theory (T), used the anomalous motion of Uranus (D) to construct a more specific theory of gravitation (T*). Yet, T* implies an independent prediction D* that there exists an undiscovered planet, and this proved to be correct when Gottfried Galle observed the posited planet—that we now call Neptune—on a telescope. Notably, in line with Worrall's idea that

[11] As Schindler noted, since actual usage is not the focus in Worrall's strong version, "the label "heuristic predictivism" no longer seems appropriate (Schindler 2014, p.64). Hence, I've opted to simply call Worrall's account "Worrall's predictivism".

D did not confirm but D* did confirm T, Adams appeared to understand that it was Galle's discovery (D*) rather than his construction of T* using D that truly confirmed T:

> I felt confident that in [the anomalous motion of Uranus], as in every previous instance of the kind, the discrepancies which had for a time thrown doubts on the truth of the law, *would eventually afford the most striking confirmation of it* (*ibid*, p.7, added emphasis).

The second category of unconditional confirmation provided by Worrall is when T* is a "'natural version'" of T (Worrall 2006, p.51). In this case, D does confirm T because it "'drops naturally out' of T" (*ibid*). What does it mean for T* to be a 'natural version' of T or for D to 'drop naturally out' of T? Worrall seemed to have in mind cases where the free parameter of a general theory is fixed according to "theoretical considerations" rather than "on the basis of the evidence" (Worrall 2005, p.819). One of Worrall's go-to examples to illustrate this is Nicolaus Copernicus' heliocentrism theory (T) and the retrograde motion of Mars (D). According to heliocentrism, both Mars and Earth are non-stationary and orbit the Sun. Accordingly, observations of Mars is made from a *moving* point of view, namely Earth. Furthermore, since Earth has a smaller orbit than Mars, it will occasionally overtake Mars' orbit and when this happens, Mars will appear to move 'backwards', i.e., retrograde. Here, the retrograde motion of Mars (D) 'drops naturally' out of heliocentrism (T).[12] Now, Worrall was well aware of Pierre Duhem's thesis that background auxiliary hypotheses are always required for a theory to entail any evidence, and he recognized that "the heliocentric hypothesis *alone* does not entail the phenomena" (Worrall 2006, p.50). Thus, D 'dropping naturally out of' T does *not* mean that T itself entails D. What does it mean then? Worrall gave no precise definition but his comment that the auxiliary hypotheses required for heliocentrism to entail the retrograde motion of Mars were "*so* direct" and "*fewer*" (*ibid*, p.49-50, original emphasis) compared to other more convoluted cases like geocentrism's epicycles does suggest that something like simplicity is what he had in mind when speaking of theoretical considerations.

In short, Worrall's predictivism distinguished conditional confirmation, where the general theory is not confirmed, and unconditional confirmation, where the general theory is confirmed. It is worth highlighting the predictivist character of Worrall's account. Notice that Worrall's distinction between conditional confirmation and unconditional confirmation can be made in terms of accommodation and prediction. In conditional confirmation, D was *used* to construct T* conditional on T. In the first type of unconditional confirmation, D was also used to construct T* but independent *predictions* (D*) are also entailed by T*. In the second type of unconditional confirmation, D was *not needed* for construction since T 'naturally

[12] It is unclear to me what the specific theory T* is supposed to be in this example. Some have understood Worrall to mean that in such examples, there *is no* T*: Schindler for instance, said "There will therefore be no [T*]" when D 'drops naturally out of' T (Schindler 2014, p.64). Worrall's writing at times suggest that there *is* some T*: "the chief question will be whether some parameter having a fixed value in [T*] was set at that value by theoretical considerations, or as a 'natural consequence' of such general considerations, or whether instead the value was fixed on the basis of the evidence" but at other times as if there is no T*: "there is no parameter within that theory that could have been fixed" (Worrall 2005, p.819). In any case, I think that this ambiguity will not affect my arguments.

entails' it.[13] Furthermore, Worrall held that unconditional confirmation was superior to conditional confirmation: "the second, unconditional and hence more *powerful*, sort of confirmation" (Worrall, 2006, p.51, added emphasis); and "[t]he *stronger* sort of confirmation that I have highlighted is the sort that spills over from the specific theory that entails the relevant data to the underlying general theory or programme" (*ibid*, p.53, added emphasis).

Accordingly, a Worrallian argument for preregistration must involve its informativeness in practically distinguishing conditional confirmation from unconditional confirmation. Since there are two types of unconditional confirmation, then there are two ways that preregistration might be informative. Firstly, for independent predictions (D*) that confirm the general theory (T), "independence" here clearly means *not used* in constructing the specific theory (T*). Take the example of Newtonian gravitation: the D* here concerns the existence of a planet disturbing Uranus' orbit, and it is entailed by T* which posits the existence of said planet. A sufficient condition for a prediction to be independent (i.e., not used) is temporal precedence: if T* temporally precedes D*, then D* *must* be independent since it could not have been used to construct T*. Accordingly, preregistration does inform us on whether T* temporally precedes D* since researchers are made to specify their hypotheses *before* data collection. Thus, preregistration allows us to practically distinguish conditional confirmation from unconditional confirmations involving independent predictions.

Secondly, for T* that is a 'natural version' of T, we need to know if T* was produced via theoretical reasons or via empirical data. As Worrall said: "the chief question will be whether some parameter having a fixed value in [T*] was set at that value by theoretical considerations, or as a 'natural consequence' of such general considerations, or whether instead the value was fixed on the basis of the evidence" (Worrall 2005, p.819). One might worry that in considering the history of the theory, specifically, *how* it was constructed, we will again require scrutinizing the psychological states of researchers and in that case, Worrall's strong version collapses back into the weak version that he took pains to deny. Worrall's response to this objection was:

> It cannot be emphasised sufficiently that 'means of construction' is, in the mature sciences at least, *not* a personal notion—finding out about it does *not* require combing through a scientist's personal diaries and the like. It depends instead on the research programmes involved. And these programmes can be articulated and objectively assessed (Worrall 2006, p.52).

That is, given a well-developed theory—a mature research program—it will be obvious whether the theory was constructed via theoretical or empirical means. Take heliocentrism and the retrograde motion of Mars, Worrall's idea there was that heliocentrism obviously entailed retrograde motion with no need of empirical data:

> the only question is whether there was some free parameter in some theory available to him which could be fixed on the basis of stations and retrogressions to produce his

[13] Consider Worrall's comment: "Copernican theory, in my view, genuinely *predicts* stations and retrogressions" (Worrall 2006, p.49, original emphasis).

> heliocentric theory (*and patently there was not*) (Worrall 2005, p.819, added emphasis).

Now, whether it is indeed the case that the means of construction are obvious for mature research programs is a question that will take me too far afield. For my purposes, I claim only that for *non-mature* research programs, it is not at all obvious whether the theory was constructed via theoretical or empirical means. If so, then scrutinizing the psychological states of researchers is back on the table and therein lies a justification for preregistration.[14] To illustrate the point concretely, consider a hypothetical example from Andrew Gelman and Eric Loken (2019, p.3): suppose that a social scientist is interested in studying whether Democrats and Republicans differ in performance in a mathematical test that either involves healthcare or military contexts. Based on relevant social science theory, the researcher hypothesizes that Democrats will perform better when the test involves healthcare while Republicans will perform better if it involves the military (presumably because of their respective familiarities with the topics). Yet, Gelman and Loken rightly point out that there remain many other choices that the researcher must make to perform a standard statistical analysis. For example, should the analysis be restricted to only males or females? A theoretical motivation for focusing on males might be that they tend to be more ideological. But another theoretical motivation for focusing on females might be that they are more sensitive to contexts. The point here is that theoretical reasoning in non-mature research programs like in much of social science is *too easy*—one gets the feeling that a 'compelling' theoretical reason can always be found. But if theoretical reasons are so readily available, then a researcher can easily claim to have constructed their theory via theoretical means *even if* it was motivated by data. It should be noted that Gelman and Eric were emphasizing that theoretical flexibility remains a big problem even if we assume that researchers do not deliberately engage in p-hacking or HARKing. But if we do include such practices—which, as mentioned above, we know are commonly practiced (John et al. 2012)—then the blurring of theoretical and empirical means in constructing theory becomes even more severe.

Worrall was certainly right in claiming that the history of a research program—specifically, how it was constructed—is a fact that can be objectively assessed, but he failed to warn us about the potentially immense practical difficulty in *finding out* this fact, especially in non-mature research programs. Worrall took pains to dissociate his account from the seemingly subjective task of scrutinizing researchers' mental states and preferred to emphasis objective historical facts about research programs. Yet, Worrall would surely not object that *if* we knew what was going on in researchers' heads, then we would know the objective facts about a research program's genesis. That is, psychological facts are a *means* to the facts concerning a research program's construction that Worrall's account is ultimately concerned with. But preregistration is precisely such a *means*. Of course, preregistrations do not literally record the thoughts of researchers, but they are informative of the decisions and their associated reasons that a researcher was making when constructing theories. More

[14] It must be emphasized here that I am not attempting to mount a counterargument against Worrall's predictivism. Rather, I am trying to explicate the consequences of Worrall's account and show how a justification for preregistration can be made. These consequences might not be favorable for Worrall's account in the context of addressing Musgrave's and Leplin's arguments, but this is not a concern for me. My aim in this part of the paper is to give reasons for why you should support preregistration *if* you are a predictivist—not that you *should* be a predictivist in the first place.

significantly, preregistration can assure Worrall that the researcher's theory was *not* constructed using empirical means since the data did not exist prior to the theory's construction. In short, since Worrall's account is committed to an epistemic difference between conditional confirmation and unconditional confirmation, and since preregistration is informative in practically distinguishing the two, then Worrall has reason to support preregistration.

3.6 Summary

In sum, I argued above that predictivists of all stripes have reason to support preregistration. This unity is especially important for two reasons. Firstly, proponents of preregistration like Nosek and colleagues appear committed to strong predictivism. The problem, however, is that strong predictivism has long been rejected by philosophers of science as a nonstarter due to various counterarguments (Howson 1990; Collins 1994; Mayo 1996; Hitchcock and Sober 2004). These counterarguments are widely held to have dealt a decisive blow to strong predictivism (Barnes 2022). Another problem is that preregistration's ability to practically distinguish predictions from disguised accommodations depends on there being minimal deviation between preregistration and manuscript—but as noted above, such deviations are common. Notice that there are *innocent* deviations such as sample size changes due to unexpected funding constraints and *nefarious* deviations resulting from disguising accommodations. But the existence of the former confounds disguised accommodations, thereby limiting preregistration's ability to distinguish predictions from disguised accommodations. Preregistration proponents can avoid these objections by retreating from strong predictivism to weak predictivism or Worrall's predictivism and still have good grounds to support preregistration.

The second reason is that the empirical efficacy of preregistration is contentious, as we have seen above. But an empirical argument for preregistration is less urgent if we at least have theoretical reasons to support preregistration. To be clear, I am not suggesting that a theoretical defense of preregistration obviates the need for empirical evidence of efficacy. The latter strikes me as a matter of willingness and ingenuity to design appropriate rules and mechanisms for preregistration to work as intended. For example, how far are we willing to go to ensure that journal reviewers cross check preregistrations against submitted manuscripts when reviewing papers? As an analogy, consider democracy and the issue of voting fraud. Suppose that on democratic principles—say, one-person-one-vote—it is argued that our elections will be more democratic if we possessed an anti-fraud machine to detect illicit extra votes. It may be entirely unclear whether the proposed anti-fraud machine will be reliable enough to work as intended, but the *reason* for why it is desirable is a separate question. What I hope to have achieved here is to establish that there are good reasons for why we should desire an analogous machine (preregistration) for science. Its empirical efficacy remains contentious, but its desirability is now clear.

Having argued that predictivists of all stripes have reason to support preregistration, I will now show that preregistration returns the favor to predictivism. As it turns out, *preregistered* predictions can be used as a counterargument to accommodationism—the antithesis of predictivism.

4. The Preregistration Argument for Predictivism

Dellsén (forthcoming, pp.9-15) recently argued that accommodations are indicative of a data fidelity advantage over predictions. "Data fidelity" refers to the extent that a set of data is free from fabrication and manipulation. "Data fabrication" refers to making up nonexistent data points or altering existing data points. "Data manipulation" refers to nontransparent reporting of results, such as reporting results of only one outcome variable when multiple was measured or removing data points without good justification. As noted above, data manipulation practices in the context of significance hypothesis testing are called "p-hacking". Data manipulation need not always be intentional, as a researcher may unintentionally distort data according to their biases. For example, the great statistician Ronald Fisher famously rejected observational evidence purporting that smoking causes lung cancer. Fisher had a principled reason for rejecting that smoking causes lung cancer, namely that the evidence did not come from a Randomized Controlled Trial (RCT) which for him was the gold standard. Nevertheless, Fisher analyzed observational data in an attempt to disprove the causal link between smoking and lung cancer (Fisher 1958). The important point to note here is that Fisher was himself a smoker and so his rejection of observational evidence could have been—at least partly—a result of motivated reasoning (Stolley 1991).

Dellsén argued that accommodations are indicative of higher data fidelity compared to predictions because researchers face little motivation to fabricate or manipulate accommodated data. There is little motivation to fabricate or manipulate accommodated data because the researcher is simply generating hypotheses based on whatever the data turns out to be. On the other hand, there is immense pressure to fabricate and manipulate predicted data because of the risk of it refuting the hypothesis. This pressure comes from journals' biases, where negative results are seen as unpublishable, and from researchers' unwillingness to be proven wrong. In sum, researchers accommodating data is less likely to intentionally or unintentionally engage in fabrication or manipulation compared to researchers predicting data, and therefore, accommodations are indicative of higher data fidelity compared to predictions.

It should be first noted that the data fidelity of accommodations may be exaggerated. The reason is that not all hypotheses are equal in the eyes of researchers. For example, a researcher will probably prefer simple, novel, unintuitive, or even "beautiful" hypotheses over complex, traditional, intuitive, or 'ugly' hypotheses. Such preferences might be due to practical considerations, as simple, novel, and unintuitive hypotheses are more likely to be published (Smaldino and McElreath 2016), or they might be due to aesthetic reasons (Schindler 2018; Ivanova 2023). Another reason why accommodation is not immune to data manipulation or fabrication is that researchers are likely to prefer accommodated hypotheses that coheres with research program rather than one that is either irrelevant or even contradicts it. All of these is just to say that it is simplistic to think that researchers have zero motivation to fabricate or manipulate accommodated data. Dellsén is likely to respond that these concerns apply equally to predictions, thus even if accommodations are not immune to data manipulation or fabrication concerns, the point remains that accommodations have *better* data fidelity than predictions. This response is well taken but I shall argue next that the data fidelity advantage of accommodations dissolves when we consider *preregistered* predictions.

Before I lay out my response, it should first be clarified that weak predictivism and weak accommodationism need not be incompatible because each position may be advocating for different epistemic advantages that can simultaneously obtain (Dellsén forthcoming, p.9). However, as we shall see, the two positions do become incompatible when the same epistemic advantage is under discussion. Recall Dellsén's claim that accommodations are indicative of higher data fidelity compared to predictions because researchers working with accommodated data have little motivation to fabricate or manipulate their data, which is in turn because they do not face pressures from publication bias and motivated reasoning. But notice that similar advantages apply to *preregistered* predictions. Data manipulation is kept in check because preregistration forces researchers to commit to (among other things) their sampling plans and statistical analyses before data is collected or analyzed. This means that the preregistered data sets are free from motivated reasoning, as long as researchers do not deviate significantly from their initial commitments. Furthermore, as noted above, Registered Reports remedies publication bias, which is one form of nontransparent reporting of results, i.e., data manipulation. Preregistration repositories also help with data fabrication as the raw data sets made available on them can be used to catch fraud.[15] For example, *Data Colada*—a blog operated by Joe Simmons, Leif Nelson, and Uri Simonsohn—recently alleged that four papers co-authored by Francesco Gino—a professor (previously) at Harvard—contains fraudulent data.[16] Importantly, the evidence that *Data Colada* consisted of anomalies in the excel files associated with the four papers (Data Colada 2023). To be sure, publishing raw data sets is not required for preregistration itself, but journals that subscribe to the Open Science movement often have separate requirements for raw data sets to be made available (e.g., PLOS One). Furthermore, journals that utilize the Registered Reports format can require raw data sets to be made available for the second stage of peer review (e.g., Nature). In sum, the data fidelity advantages of accommodations identified by Dellsén similarly applies to preregistered predictions too.

To be sure, I am here slightly redefining the predictivism debate to be about preregistered predictions rather than simply predictions, but this redefinition is warranted given that the scientific community is starting to adopt the view that preregistration is necessary for legitimate prediction. Relatedly, the history of randomization is another example of evolving norms in experimental design. Hacking (1988) noted that randomization is so commonplace then—and now—that we forget that its popularity was not always the case. In lieu of randomization, Gosset and others advocated for 'matched' assignment designs during the early 20th century (*ibid*, p.429). Interestingly, we can trace origins of both randomization and preregistration to parapsychology research. For randomization, Richets was a pioneer in employing randomization in parapsychology in the 19th century (pp.437-440). For preregistration, the *European Journal of Parapsychology* (EJP) was the first psychology journal to accept and publish preregistered research (Wiseman et al. 2019). Besides similar origins in parapsychology, there is also evidence that preregistration is following randomization in popularity. Consider, for example, that more than 300 journals now accept

[15] A reviewer noted that the anti-fraud here is caused by open data rather than preregistration. It is worth noting that open data and preregistration often go hand-in-hand because journals that require the latter also often require the former. Thus, the reviewer's point is well-taken that preregistration *correlates* but does not cause anti-fraud. But correlation is all I need since Dellsén and I are concerned only with whether prediction and accommodation *indicate* (not cause) high data fidelity.

[16] What is said here should not be taken as endorsement (nor rejection) of *Data Colada*'s claims.

manuscripts submitted using the Registered Reports format (Chambers and Tzavella 2022). Further evidence of evolving norms can be seen from the fact that some psychologists have openly called for preregistration to be a necessary condition for genuine prediction (Wagenmakers et al. 2012). If preregistration is becoming the norm, then it is entirely legitimate to compare preregistered predictions against accommodations.

5. Conclusion

I argued above that there are interesting connections between the new focus in the sciences on preregistration as a solution to replication crises and the older predictivism debate in the philosophy of science. On the one hand, the predictivist thesis that predictions are epistemically superior to accommodations explains why preregistration proponents desire to distinguish predictions from accommodations. Strong predictivism gives the most straightforward justification for preregistration, namely that it is a tool to distinguish between confirmatory predictions from non-confirmatory accommodations. The weak predictivist justification for preregistration is less straightforward since for them the prediction-accommodation distinction is consequential only insofar as it tracks other epistemic properties. Despite these seemingly disparate epistemic properties, I argued that weak predictivists of all stripes and Worrall's predictivism should support preregistration as it informs them about the epistemic properties that they deem important. On the other hand, I argued that Dellsén's argument for weak accommodationism dissolves once we consider preregistered predictions rather than predictions *per se*. In short, predictivism justifies preregistration and preregistration returns the favor by bolstering the predictivist thesis.

References

Adams, W.A. (1896). *The Scientific Papers of John Couch Adams, Vol I*. Cambridge University Press, Cambridge.

Bak-Coleman, J., & Devezer, B. (2024). Claims about scientific rigour require rigour. *Nature Human Behaviour*, *8*(10), 1890–1891. https://doi.org/10.1038/s41562-024-01982-w

Barnes, E. C. (2022). "Prediction versus Accommodation", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/win2022/entries/prediction-accommodation/

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Chambers, C. D., & Tzavella, L. (2021). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), 211037. https://doi.org/10.1098/rsos.211037

Collins, R. (1994). Against the Epistemic Value of Prediction Over Accommodation. *Noûs*, *28*(2), 210. https://doi.org/10.2307/2216049

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai Van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., … Zhou, X. (2021). Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*, *12*(1), 9–44. https://doi.org/10.1007/s13164-018-0400-9

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00621

Ensinck, E. N. F., & Lakens, D. (2023). *An Inception Cohort Study Quantifying How Many Registered Studies are Published*. https://doi.org/10.31234/osf.io/5hkjz

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

Fife, D. A., & Rodgers, J. L. (2022). Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the "replication crisis". *American Psychologist*, *77*(3), 453–466. https://doi.org/10.1037/amp0000886

Fisher, R. A. (1958). Cancer and Smoking. *Nature*, *182*(4635), 596–596. https://doi.org/10.1038/182596a0

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Gelman, A., & Loken, E. (2019). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time.

Giere, R. N. (1984). *Understanding scientific reasoning* (2nd ed). Holt, Rinehart, and Winston.

Heirene, R., LaPlante, D., Louderback, E., Keen, B., Bakker, M., Serafimovska, A., & Gainsbury, S. (2024). Preregistration specificity and adherence: A review of preregistered gambling studies and cross-disciplinary comparison. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2021.2909

Hitchcock, C., & Sober, E. (2004). Prediction Versus Accommodation and the Risk of Overfitting. *The British Journal for the Philosophy of Science*, *55*(1), 1–34. https://doi.org/10.1093/bjps/55.1.1

Howson, C. (1990). *Fitting Your Theory to the Facts: Probably Not Such a Bad Thing after All*. In: Savage, C.W., Ed., Minnesota Studies in the Philosophy of Science, Vol. 14, University of Minnesota Press, Minneapolis, 224-244.

Ivanova, M. (2023). What is a Beautiful Experiment? *Erkenntnis*, *88*(8), 3419–3437. https://doi.org/10.1007/s10670-021-00509-3

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Lakens D. (2019). *The value of preregistration for psychological science: A conceptual analysis* (No. 3). 心理学評論刊行会. https://doi.org/10.24602/sjpr.62.3_221

Lakens, D. (2023). *Concerns about Replicability, Theorizing, Applicability, Generalizability, and Methodology across Two Crises in Social Psychology*. https://doi.org/10.31234/osf.io/dtvs7

Lakens, D., Mesquida, C., Rasti, S., & Ditroilo, M. (2024). The benefits of preregistration and Registered Reports. *Evidence-Based Toxicology*, *2*(1), 2376046. https://doi.org/10.1080/2833373X.2024.2376046

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton Univ. Press.

Longino, H. E. (2002). *The fate of knowledge*. Princeton University Press.

Maher, P. (1988). Prediction, Accommodation, and the Logic of Discovery. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1988*(1), 272–285. https://doi.org/10.1086/psaprocbienmeetp.1988.1.192994

Mayo, D. G. (1991). Novel Evidence and Severe Tests. *Philosophy of Science*, *58*(4), 523–552. https://doi.org/10.1086/289639

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Protzko, J., Krosnick, J., Nelson, L., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O’Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., & Schooler, J. W. (2023). RETRACTED ARTICLE: High

replicability of newly discovered social-behavioural findings is achievable. *Nature Human Behaviour*, *8*(2), 311–319. https://doi.org/10.1038/s41562-023-01749-9

Protzko, J., Krosnick, J., Nelson, L., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., & Schooler, J. W. (2024). Retraction Note: High replicability of newly discovered social-behavioural findings is achievable. *Nature Human Behaviour*, *8*(10), 2067–2067. https://doi.org/10.1038/s41562-024-01997-3

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 251524592110074. https://doi.org/10.1177/25152459211007467

Schindler, S. (2014). Novelty, coherence, and Mendeleev's periodic table. *Studies in History and Philosophy of Science Part A*, *45*, 62–69. https://doi.org/10.1016/j.shpsa.2013.10.007

Schindler, S. (2018). *Theoretical virtues in science: Uncovering reality through theory*. Cambridge University Press.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Stolley, P. D. (1991). When Genius Errs: R. A. Fisher and the Lung Cancer Controversy. *American Journal of Epidemiology*, *133*(5), 416–425. https://doi.org/10.1093/oxfordjournals.aje.a115904

Szollosi, A., & Donkin, C. (2021). Arrested Theory Development: The Misguided Distinction Between Exploratory and Confirmatory Research. *Perspectives on Psychological Science*, *16*(4), 717–724. https://doi.org/10.1177/1745691620966796

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., Van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Van Den Akker, O. R., Van Assen, M. A. L. M., Bakker, M., Elsherif, M., Wong, T. K., & Wicherts, J. M. (2023). Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02277-0

Van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Van Der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wiseman, R., Watt, C., & Kornbrot, D. (2019). Registered reports: An early example and analysis. *PeerJ*, *7*, e6232. https://doi.org/10.7717/peerj.6232

Worrall, J. (1985). Scientific Discovery and Theory-Confirmation. In J. C. Pitt (Ed.), *Change and Progress in Modern Science* (pp. 301–331). Springer Netherlands. https://doi.org/10.1007/978-94-009-6525-6_11

Worrall, J. (2005). Prediction and the 'periodic law': A rejoinder to Barnes. *Studies in History and Philosophy of Science Part A*, *36*(4), 817–826. https://doi.org/10.1016/j.shpsa.2005.08.007

Worrall, J. (2006). Theory-Confirmation and History. In C. Cheyne & J. Worrall (Eds.), *Rationality and Reality* (pp. 31–61). Springer Netherlands. https://doi.org/10.1007/1-4020-4207-8_4

Yamada, Y. (2018). How to Crack Pre-registration: Toward Transparent and Open Science. *Frontiers in Psychology*, *9*, 1831. https://doi.org/10.3389/fpsyg.2018.01831