Falsificationism – the incomplete story about choosing good experiments

Rafael Fuchs (Rafael.Fuchs@campus.lmu.de)[1] **& Ulrike Hahn (u.hahn@bbk.ac.uk)**[2]
[1]Munich Center for Mathematical Philosophy and Graduate School of Systemic Neurosciences, LMU Munich , 80539 Munich, Germany
[2]Centre for Cognition, Computation, and Modelling, Birkbeck, University of London , London, WC1E 7HX, U.K.

Abstract

Falsificationism remains a popular framework considered by many scientists as 'the normatively correct model of science', and of human cognition as a kind of 'lay science'. In this paper, we demonstrate some fundamental conceptual shortcomings and gaps in this normative prescription. While some of the issues presented here connect to well-known arguments from the philosophy of science, we connect these issues to the topic of efficient information search. Specifically, we show that a strict falsificationist rejection of the idea of 'confirmation' of scientific hypothesis (more specifically, entertaining and updating beliefs in those hypotheses, based on evidence) makes scientific research and information search overall less efficient, compared to an accuracy-based Bayesian approach.

Keywords:

Information Search; Metascience; Optimal Experimental Design; Falsificationism; Bayesian Philosophy of Science;

Introduction

Falsificationism remains a popular framework that continues to be seen by many scientists as 'the normatively correct model of science' (Lakens, 2024; Lakens et al., 2018). Although its status within the philosophy of science has, arguably, waned (for examples, see e.g., Howson and Urbach, 2006; Sprenger and Hartmann, 2019) it continues to shape methodological choices in empirical sciences such as the choice between frequentist and Bayesian statistics (Wagenmakers et al., 2008). This gives it a twofold relevance to cognitive science. It continues to influence the design of empirical studies and the evaluation of their results (in the form of 'methodological falsificationism', Lakatos et al., 1978). However, it also remains relevant in contexts where we seek to construe human cognition as a kind of 'lay science'. Not only are there strands of research that explicitly draw on the model of science's truth seeking process, such as the idea of the developing child as a scientist (Gopnik, 1996; Karmiloff-Smith, 1988; Kuhn, 1989), there are many areas concerned with the human acquisition of knowledge that are influenced, directly or indirectly, by debates in epistemology and the philosophy of science. Last but not least, fundamental ideas in those disciplines continue to influence our construal of the task humans face in trying to come to a (more or less) accurate understanding of the world. Falsificationism is one of those ideas.

There have been many different critiques of falsificationism, and even reviewing them would be beyond the scope of this paper. The topic of the current paper is an issue that has been touched upon, but, to the best of our knowledge, has not been explicitly developed in this way in prior research: specifically, it is the way in which falsificationism connects the search for evidence to our current beliefs. It is typically seen as problematic to systematically prefer searching for evidence that, if found, would confirm our beliefs. Different manifestations of this underlying intuition are studied under the header of confirmation bias (Nickerson, 1998). Falsificationism, however, seemingly not only compels us to even-handedly seek out what would just be strong evidence, regardless of whether that evidence would be evidence for or against our currently favoured, underlying hypothesis, it asks us specifically to seek out evidence *against* a hypothesis, even (or particularly) when we prefer it and consider it, more likely than not, to be true.

It has been demonstrated clearly before that searching for dis-confirming evidence is not always more informative than seeking for confirmation (see within psychology, for example, Klayman and Ha, 1987; Oaksford and Chater, 1994). However, what has not (to the best of our knowledge) been in focus is the fact that the pursuit of falsification is frequently, or even typically, *counter-attitudinal*. Falsification seems to compel us to differentially seek out evidence to refute our currently preferred theory. It thus not only decouples the search for evidence from our current degree of belief in a hypothesis, in practice it often effectively mandates us to go against it. Where we currently hold a theory likely to be true, it is, in effect, asking us to seek out evidence that we think likely doesn't exist, namely evidence that would refute that theory.

While it is open to question whether cognitive scientists actively seek to falsify their theories with their experiments, it nevertheless remains the case that they are frequently told that this is what they ought to do (see e.g., discussion in Holtz and Monnerjahn, 2017). This gives the ideal of falsification continued relevance. In this paper, we explore the implications of that prescription. After presenting a short review of falsificationist core ideas for scientific inquiry, we discuss some conceptual gaps and shortcomings with respect to effective experiment selection, evidence strength, and 'risky predictions'. Some of these arguments are already well-known, but here we extend and connect them to the issue of efficient information search. Specifically, we show that a strict falsificationist rejection of the idea of 'confirmation' of scientific hypothesis (more specifically, entertaining and updating beliefs in those hypotheses, based on evidence) makes scientific

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

research and information search overall less efficient, compared to an accuracy-based Bayesian approach.

Background: Falsificationist Core Ideas

Falsificationism gets its name from the observation that there is an asymmetry between 'verification' (or 'confirmation') of scientific theories and *falsification*. The idea is based on a simple logical fact. A scientific theory is a generalisation over a large (possibly infinite) domain, i.e., a universally quantified statement. To 'verify' a universally quantified statement would mean verifying individual instances over the whole domain, which cannot be achieved in finite time (especially since theories are intended to work for future instances as well, which may or may not behave according to the current theory – this is just the well-known problem of induction). On the other hand, *falsifying* a universal statement only requires a single counterexample. Therefore, Popper's criterion of demarcation for scientific theories, and the rational process of science (which is supposed to sidestep the problem of induction) is centred around *falsification*: given a general theory, scientists deduce predictions for singular, concrete instances, and test these predictions experimentally. If the actual observation contradicts the prediction, the theory is refuted, so we need to come up with a new one that works better – otherwise, keep working with the theory tentatively (in particular, keep testing it until refuted, if ever). This is the process that Popper termed 'conjectures and refutations' (come up with a theory, test it until refuted, repeat), which he considered to be the core of the scientific process and progress. This process is intended to be fully deductive (come up with some hypothesis, and deduce implications for testing), by which Popper wanted to eliminate the problem of induction by eliminating the notion of induction itself ('induction is a myth'). Importantly, a strict reading of this position means that there is no such thing as (partial) verification or confirmation of theories. Whenever the actual observations fit the predictions of the theory, this only means that the theory is *not falsified*, rather than *positively confirmed*. Popper was also very adamant that scientists should choose experiments with the goal of trying to falsify their theory. He famously wrote that finding 'confirmatory instances is easy' – we can always find some instance that is consistent with our theory, or make the theory more flexible so that it can cover *any* instance (in the extreme case, the theories 'predictions' are tautological, which makes no informative claim about the world at all). The hallmark of a good theory (and correspondingly, a good experiment to test the theory) is a *risky prediction*.

Evidence Strength, 'Risky Predictions', and Plausible Evidence Distributions

There are many important insights that are articulated in Popper's writing, such as the idea that it is not a virtue of an empirical theory if it is so malleable that it can accommodate virtually all data, but rather a problem. It is most definitely not a desirable feature of an empirical hypothesis that such flexibility makes it *unfalsifiable*. Nevertheless, it is a corollary of an empirical claim actually being true that it won't be possible to falsify it in practice.

This fact is embodied in the widespread intuition (both among scientists and statisticians) that, although methodological falsificationism cannot directly confirm a hypothesis, it will nevertheless essentially do so in the long run by virtue of the fact that our rigorously tested hypothesis is 'the last man standing'.

The core conceptual limitation underlying Popper's foundation for falsificationism, however, is its failure to reckon appropriately with the fact that evidence *varies in strength*, that is, its diagnosticity is a matter of degree. It is an inevitable consequence of the logical positivist outlook that any 'evidence' that does not refute a theory in some ways 'confirms it' and that seems to lie at the core of Popper's perception that finding confirming evidence is 'easy'. However, what is patently possible in our everyday experience, both as lay reasoners and as scientist, is finding *strong* evidence that is confirming.

This is well-illustrated with the so-called Raven's Paradox, which arises from the idea that observations corresponding to logically equivalent formulations of a theory are treated as equally confirmatory (e.g. 'all ravens are black' is equivalent to 'all non-black objects are non-ravens', and therefore observing a white shoe or a red chair confirms the hypothesis that all ravens are black just as much as observing a black raven does). The Bayesian framework (Howson & Urbach, 2006), by contrast, treats this as a sampling problem, which takes into account that the *number* of objects that are 'non-ravens' is several orders of magnitude higher than the number of ravens, and therefore observing a white shoe is almost negligible evidence for the hypothesis that all ravens are black.

The fact that evidence fundamentally varies in strength, whether it be evidence for or against a hypothesis, has not just featured in multiple ways in Bayesian arguments against falsificationism (e.g. Jeffrey, 1975), and, along with it, frequentist statistics (more on that at the end of the section), making methodological falsification workable in practices has required introducing that notion. In line with Popper's notion of 'risky predictions', contemporary accounts of methodological falsificationism stress the importance of so-called *severe tests* (Lakens, 2024; Mayo, 2018). A severe test is one where:

> 1) it is likely that a test will support a hypothesis when the hypothesis is correct, and 2) it is likely that a test will fail to support a hypothesis when the hypothesis is wrong (Lakens, 2024, pg. 1)

Conceptually, those familiar with signal detection theory (Green, Swets, et al., 1966) will recognise these two quantities as the sensitivity of a test (1) and its specificity (2). In the context of Bayes' theorem, they correspond to $P(e|H)$ (1) and $P(\neg e|\neg H)$ both being high.[1]. Because $P(\neg e|\neg H)$

[1]Bayes' Theorem states $P(H \mid e) = \frac{P(e|H) \cdot P(H)}{P(e)}$, which can also be represented as a relation between odds- and likelihood ratios:

constrains the 'false positive rate' of a test[2] these two quantities also determine the so-called Likelihood Ratio (LHR), $P(e|H)/P(e|\neg H)$, which is widely used as a measure of diagnosticity, that is of evidence strength (Fitelson, 1999; Nordgaard & Rasmusson, 2012).

What falsificationism overlooks, however, is that not only is the strength of evidence connected to the probability of finding it, but that the probability of finding evidence for or against the hypothesis is inherently connected to whether or not it that hypothesis is true. By the same token, our rational expectations of obtaining that evidence are connected, from a Bayesian perspective, to our current beliefs in the hypothesis.

That strong (i.e. diagnostic) evidence is less probable is a well-known fact that follows from what is known as total probability (see also Jeffrey, 1975):

$$P(e) = P(e|H)P(H) + P(e|\neg H)P(\neg H) \quad (1)$$

From the fact that this constitutes the denominator of Bayes' rule, it follows also that less likely, more 'surprising' evidence, leads to greater change in belief (and with that to greater confirmation, though that relationship is complicated by the fact that there are numerous competing notions of confirmation, see e.g., (Crupi, 2025; Fitelson, 1999).

However, it also follows from total probability that positive evidence (i.e., evidence that supports the hypothesis) is more likely than negative evidence *if the hypothesis is true*. By the same token, the difference between the chance of encountering positive evidence (e) vs. negative evidence ($\neg e$) changes, in expectation, as a function of our current degree of belief in hypothesis H (for details see Hahn, 2025).

All of these facts follow directly from probability theory given the construal of evidence strength in terms of the likelihood ratio, and with that the construal in terms of severity.

This is problematic for falsificationism in several ways. First, it identifies the problem with one of the driving intuitions underlying falsificationism, namely that confirmatory evidence is useless because 'it can always be found'. Weak evidence is easier to find than strong evidence, but that holds for both confirming and dis-confirming evidence. And whether or not confirming evidence is easier or harder to find (in the specific sense of 'more probable'), depends on whether or not H is true. By contrast, whether evidence is 'confirming' or 'dis-confirming' with respect to the hypothesis, on its own, is *entirely irrelevant* to how easy or hard evidence is to find.

Second, given that evidence strength clearly matters, both from the perspective of methodological falsificationism and Bayesianism, the conceptual and practical limitation surrounding severity is consequential. Crucially, the notion as defined leaves an important question entirely open: the two aspects of severity ([1] i.e., $P(e|H)$) and ([2], i.e., $P(\neg e|\neg H)$) are independent. They can, and routinely do, come apart in real world tests. Test sensitivity can be high or low; what matters for the LHR, and with that the diagnostic value of the evidence is the *relationship* between sensitivity and specificity. The notion of severity leaves entirely open how to choose between tests that vary with respect to the two quantities, whereas thinking about this in terms of the LHR (from a Bayesian framework) makes clear that we should simply prefer (all other things equal) the more diagnostic test. Here is an example that illustrates this problem (adapted from Cohen, 1994). Consider the following inference:

1. Less than 1 percent of Americans are members of congress;
2. This (randomly chosen) person *is* a member of congress.
3. Therefore, we reject the hypothesis that this person is an American.

What went wrong here? Clearly, we forgot about the alternative (*only* Americans are members of congress), under which the observed evidence is even less likely (*impossible*, to be precise), and therefore the data actually *proves* the hypothesis that this person is American. This problem arises when we try to transfer the logic of falsificationism (Modus tollens, which is a valid inference in a strictly deductive setting) to a probabilistic context, which is often advocated as a theoretical underpinning of classical tests of significance ($P(e \mid H)$ is low, so from observing e we infer $\neg H$). In the probabilistic setting, the inference becomes invalid – we must take into account the alternative (which makes a strictly Popperian logic hard to apply, since we hardly ever encounter any instances of 'pure' modus tollens – in science, there is always *some* noise and uncertainty, which means that statistical reasoning is required).

Third, and finally, the more accurate perspective on what evidence is, and is not, actually 'easy to find' which is provided by probability theory brings into focus the strangeness of ignoring entirely one's degree of belief in the hypothesis in identifying suitable tests.

The very argument for falsification given by Popper is motivated by an unduly crude binary notion of confirmation that gives rise to the sense that confirmatory evidence, by virtue of its ubiquity, is worthless. It thus, in effect, draws on the probability of obtaining particular kinds of evidence to determine the meta-strategy for testing. However, by basic probability theory, the chance of obtaining particular kinds of evidence varies with the truth or falsity of the test; that is, the (statistical) expectation is mathematically dependent on the probability of the hypothesis being true.

Ignoring that information, by design, not only doesn't seem normative; it is in outright conflict with the standard construals of rationality that decision theory, provides. Decision theory compels us to maximise expected value. Failure to do so incurs loss. It should thus not be surprising that selecting tests independently of the probability of the hypothesis is inefficient. We outline this in more detail next.

$\frac{P(H|e)}{P(\neg H|e)} = \frac{P(e|H)}{P(e|\neg H)} \cdot \frac{P(H)}{P(\neg H)}$. According to Bayes' rule, this expresses the posterior belief $P(H \mid e)$ attributed to hypothesis H after having observed evidence e.

[2]The false positive rate $P(e|\neg H)$ = 1 - $P(\neg e|\neg H)$.

Efficient Experimentation Over Time

There is a vast literature on optimal experimental design and efficient information search (Crupi et al., 2018; Good, 1966; Lindley, 1956; Meder et al., 2022; Rainforth et al., 2024). One standard formula for choosing an experiment E with the highest expected information gain with respect to a given set of alternative hypotheses H is to choose E that maximises *mutual information*, i.e.

$$\begin{aligned} I(H;E) &= I(H) - I(H \mid E) \\ &= I(E) - I(E \mid H) \end{aligned} \tag{2}$$

where I is an information- or entropy measure, like Shannon entropy $I(X) = -\sum_x P(x) \cdot \log P(x)$, with the conditional expected value $I(X \mid Y) = \sum_y P(y) \cdot I(X \mid Y = y)$ (for an overview of alternative measures, see e.g. Crupi et al., 2018; Gneiting and Raftery, 2007). Interestingly, in virtue of Shannon's source coding theorem (Shannon, 1948), Shannon entropy is also a precise lower bound on the expected number of questions or experiments needed to identify the true outcome ω^* from a set of alternatives Ω given a probability distribution P. It is the expected depth of a tree that represents a uniquely decodable encoding scheme of Ω that minimises the expected code-length, corresponding to the expected number of questions that need to be asked to identify ω^*. Thus, Shannon entropy as an information measure is directly related to the question of efficient experimentation[3].

Information measures like Shannon entropy are also directly connected to the notion of accuracy, as measured by scoring rules. These scoring rules are widely employed in forecasting (Gneiting & Raftery, 2007) and formal epistemology (Levinstein, 2017; Pettigrew, 2016). Importantly, it has been shown (Rosenkrantz, 1992) that every *strictly proper* scoring rule[4] vindicates Bayesian conditionalisation, in the sense that the posterior belief obtained by Bayes' theorem minimises expected inaccuracy given the prior. Moreover, in line with that it is also known that inaccuracy over time tends to zero, since updating probabilities by conditionalisation leads to convergence on the true hypothesis, if the true hypothesis is contained in the set of alternatives (Le Cam, 1953; Gaifman and Snir, 1982; Blackwell and Dubins, 1962; Hawthorne, 1994). Importantly, this holds *even if* the likelihoods $P(e|h_i)$ are imprecise (Hahn et al., 2018), as long as they *qualitatively* 'point in the right direction' (i.e. if e is evidence *specifically for* h_i, then $P(e|h_i) > P(e|h_j)$ for all $j \neq i$), which simply means that the agent has correctly spelled out the (qualitative or quantitative) implications of every h_i. This means that subjective ('arbitrary') priors 'wash out' over time, but on top of that, scoring rules can also be used to define decision rules for objective prior assignments that are optimised for avoiding specific risks of inaccuracy from the start (e.g. the prior that maximises Shannon entropy subject to given constraints – also known as the *MaxEnt* assignment (Jaynes, 1968; Landes & Williamson, 2013; Williamson, 2010) – minimises worst-case expected inaccuracy).

For a given prior $P(h_i)$ for $h_i \in H$, eq. (2) amounts to minimising the expected posterior uncertainty $I(H \mid E)$, or equivalently $I(E \mid H)$. Now, we will examine how priors (and their trajectory over time, in virtue of truth-convergence) play an essential role for selecting informative experiments in the sense of eq. (2). As mentioned above, the central precondition of Bayesian truth-convergence is that the true hypothesis is contained in the set of alternatives. Before considering what happens if this condition is violated, let us start with the simple case, where we have a set of hypotheses $H = \{h_1, \ldots, h_n\}$ that contains the true theory h^* (e.g. trivially, if H is a partition of all possible worlds). Already at this point, there is an interesting discrepancy between Bayesianism on the one hand, and falsificationism and classical statistics on the other hand. Since the latter don't assign credences to hypotheses, there is no overall expectation of the evidence (i.e. $P(e)$, as defined in eq. (1) is undefined in the classical/falsificationist framework). Consequently, it is more difficult to entertain and test multiple alternative hypotheses simultaneously. Falsificationism (and classical statistics) usually focuses on a single hypothesis, whereas Bayesianism allows for entertaining (and investigating) multiple hypotheses simultaneously, by aggregating expectations via prior probabilities. At best, classical statistics can compare one main- and one alternative hypothesis at the same time (like in the Neyman-Pearson framework). Comparing *another* alternative hypothesis requires a new experiment, in order to control the type I error (this is known as the problem of multiple comparisons).

This is a first observation in the static setting, for choosing a single experiment to distinguish multiple hypotheses. If priors are already reasonably accurate, the experiment that maximises mutual information (eq. (2)) is optimal on average. On the other hand, one might object that initial priors may contain some inaccuracy (a common criticism of Bayesian statistics), so the more important question is what happens *over time*, as Bayesians learn and become increasingly accurate. The real punchline is that, as Bayesian posteriors become more accurate, their experiments become more efficient as well – which is something that opponents of Bayesian belief updates cannot achieve. As mentioned, prior probabilities over H are updated according to Bayes' theorem, and the resulting posteriors become increasingly accurate over time, i.e. the posterior mass increasingly concentrates around the true h^*. Moreover, by total probability (eq. (1)), this also means that our predictions within the domain of h^* become increasingly accurate, so we can make more specific predictions with in-

[3] as a side note, maximising *next-step* mutual information $I(H;E)$ will not necessarily lead to a sequence of experiments that is maximally efficient *overall*, depending on interactions between successive experiments. The Shannon entropy of an ensemble $\langle \Omega, P \rangle$ gives the expected number of experiments of the best strategy (within 1 bit of the expected value), but in order to *find* this strategy, other algorithms – e.g. like Huffman-coding (Huffman, 1952), in simple cases – are needed.

[4] a scoring rule I is strictly proper iff the expected inaccuracy $\mathbb{E}_Q[I(P)]$ of P under Q is minimal iff $P = Q$.

creasing precision. This has important repercussion for the efficient design of *future* experiments, particularly concerning a different set hypotheses H' in a logically independent domain (i.e. $h'_j \cap h^* \neq \emptyset$ for $h'_j \in H'$), where either (i) h^* figures as an *auxiliary* hypothesis, or (ii) it directly conveys information about some of the alternatives in H'.

To illustrate this, suppose a doctor is seeing a patient whose symptoms are consistent with a couple of conditions (hypothesis set H concerning the single patient). If epidemiological data indicates that there is one particular condition that is by far the most common among those alternatives (related hypothesis set H' about the population distribution), it would be inefficient to try testing for the *other* (less likely) alternatives first – naturally, we would start with a test that is specific for *that* condition[5]. Or take another example: if I realise that I forgot my wallet while going out, and I am almost certain that it is in the other jacket (from yesterday) that I left at home, then I will start looking there, rather than in the office or at my friend's place or some other random place[6] (arguably, we all do this quite naturally in everyday life, and efficient information search in science is not different in principle).

In the context of more theoretical scientific research, the transfer of information from one well-confirmed hypothesis h' to another, new (more uncertain) hypothesis set H appears in the form of auxiliary hypotheses. Suppose, we want to investigate a new set of hypotheses H (containing the true h^*), but in order to perform and evaluate any experiment E, we also need to account for possible auxiliary hypotheses A (which describe different initial conditions or auxiliary theories[7]). If H and A are independent, and both influence the prediction of outcomes in $E = \{e_1, \ldots, e_n\}$, this can be graphically represented as $H \rightarrow E \leftarrow A$ (and we can do this for every possible experiment E_i that we could perform). Equipping this directed graph with conditional probability distributions gives us a Bayesian network. The expected posterior inaccuracy after performing experiment E is given by

$$I(H \mid E, A) = \sum_{a \in A} P(a) \cdot \sum_{e \in E} P(e \mid a) \cdot I(H \mid E = e, A = a), \tag{3}$$

where (by total probability and independence) $P(e \mid a) = \sum_{h \in H} P(e \mid a, h) \cdot P(h)$. Now, we can think of different configurations of A as possible noise distributions over the relation between H and the observed experimental outcome in E. If we already have a very good idea about which $a \in A$ is actual, we don't have to take this uncertainty into account in our experimental design, and thus test the predictions of different $h \in H$ more directly and efficiently.

If the auxiliary variable A can take on different values that modify the predictions of H with respect to E in different (possibly opposite) ways, and we have high uncertainty about the actual value of A, the expectation over E becomes an average of all these possibilities, and the respective experimental result will be less informative about H than when we have information about the state of A (after all, results that diverge from theoretical expectations specified by H can be explained away by different initial conditions, e.g. broken measurement devices, and so on). In this case, a radical falsificationist who rejects the idea of positive confirmation has to separately investigate the state of A (e.g. perform calibration experiments with devices) to rule out alternatives *before* being able to test H specifically. The data collected in the preliminary experiments might also have some relation to H, but insofar as they cannot be used to rule out any of the alternatives within H (after all, they have been used to rule out alternatives within A), they cannot be used to make inferences about H[8]. So, we need more experiments. From a Bayesian perspective, on the other hand, testing can be more 'holistic': we can make positive (albeit small) inferences about H already from our preliminary data, e.g. if we find that the reliability of the measurement device was confirmed. This has important implication for the efficiency of meta-analyses and aggregation of small studies, where uncertainty about methodological reliability often leads to the exclusion of studies even with possibly-minor methodological deficits.

Thus, given that Bayesian probabilities become more accurate with increasing data samples – provided that the truth is within the set of alternatives – experimental choices and predictions based on updated probabilities become more efficient and accurate over time. Rejecting this kind of updating in the sense of strict falsificationism, taken at face value, would make science less efficient.

But what if the true h^* is not in the set of alternatives? This case is interesting, because it leads to the question of rational discovery, for which Popper explicitly denied the existence of a principled answer. We start with some set of alternatives, test them until rejection, and then come up with new alternatives – when and how we do that is not subject to any rationality constraints (we can literally 'dream up' new hypotheses at any point, as long as they are consistent with the data we already have). In the Bayesian framework, the process of introducing new theories is treated under the labels of *language change* and *awareness extension* (Steele & Stefánsson,

[5] this can also be illustrated by a simple number guessing game: suppose a random number generator selects some number between 1 and 64, and you have to guess it using only yes/no questions. If the device's choice is perfectly random, the optimal strategy is successive 'splitting in half' (i.e. $n \leq 32$? and so on). But suppose you obtain information to the effect that the generator is actually biased, and chooses either 63 or 64 half of the time. Then, it would be much more efficient to directly ask $n \geq 63$?, just as testing for the most likely condition is in the doctor-example.

[6] of course, assuming that the cost of checking each of these places (e.g. in terms of time) is roughly the same. The more general formula for choosing experiments efficiently also takes into account the expected *cost* of an experiment (so, it optimises the *ratio* between information gain (eq. (2)) and expected cost).

[7] a stock example from the philosophy of science is astronomy: H is a set of hypotheses about the orbits of certain planets, and A is the theory of optics required for operating telescopes.

[8] again, in a setting of classical statistics, this also relates to type I error control

2021; Wenmackers & Romeijn, 2016; Williamson, 2003). Generally speaking, this concerns the question of how rational agents should re-assign their probabilities when the set of alternatives changes (e.g. they become aware of new options, develop new theories and experimental techniques, or drop old terms like 'Phlogiston'). As Williamson (2003) notes, this question of language change is related to the question of prior assignments. If we have a rule for prior assignments (like classical MaxEnt or some other accuracy-based rule), we also have a rule for re-assigning probabilities after language change (first, recompute the priors over the new language, then recompute the updates on the data observed so far). It has been shown (Fuchs & Hartmann, 2024) that bounded agents who follow this procedure can emulate the probabilities of 'ideal agents' who entertain a large set of theories that includes the truth (in doing so, this procedure also eliminates the so-called 'problem of old evidence', see Glymour, 2016). Thus, Bayesian agents who follow the accuracy-based approach have a good chance of recovering truth-convergence, even if they start with a set of alternatives that might not contain the truth. Moreover, keeping track of our theories' predictive accuracy also provides some guidance and constraints for when and how to introduce new theories. Specifically, we can measure the average predictive accuracy of each theory's likelihoods, $I(P(e \mid h_i), e)$ (relative to the observed outcomes). The more the performance of *all* theories falls behind that of a given benchmark[9], the 'more urgently' a new alternative is needed. Furthermore, accuracy-based prior assignments can be used to guide the process of theory construction as well, at least to some extent. Based on a given rule for prior assignments and given data, we can try to fit our theory such that the posterior given the known data is maximal, or at least reasonably high. Importantly, rules like MaxEnt guard against overfitting, as shown in (Fuchs & Hartmann, 2024), because the prior $P(h_i)$ is a function of the conditional entropies $I(E \mid h_i)$ in relation to that of the other hypotheses $I(E \mid h_j)$ (which means that hypotheses that make extremely specific 'predictions' – as tailored to idiosyncratic data – automatically receive less prior mass, until they are confirmed by additional data).

Even if there is still work to be done, in order to perfect the accuracy-based framework, we already get an idea that the 'logic of discovery' is not as alleged by falsificationism, and consequently, the ignoring the confirmatory impact of old data also leads to efficiency loss in scientific inquiry.

[9] for example, one straightforward benchmark is a uniform random predictor (that predicts every experimental outcome with uniform probability). This is also the solution for the likelihood $P(\cdot \mid \neg h)$ of the so-called 'catch-all' hypothesis $\neg h$ under most accuracy-based rules, like MaxEnt. The catch-all hypothesis is a non-specific hypothesis, which is simply the negation of the disjunction of all explicitly formulated hypotheses. The MaxEnt solution for $P(\cdot \mid \neg h)$ is a uniform distribution, precisely because the catch-all makes no specific positive claim, i.e. there is no information on $P(\cdot \mid \neg h)$ whatsoever. If we adopt this as our benchmark, then increasing convergence on the catch-all can also be used to indicate 'how urgently' a new set of alternative hypotheses is needed.

Conclusion

In this paper, we have shown why a foundational assumption for Popper about the prevalence of confirmatory evidence is at odds with probability theory. There is nothing special, in and of itself, about 'confirmatory evidence' once one moves away from mere logical consistency to a varied notion of strength. From that same shift, however, it also follows that one should, rationally, take the probability of the hypothesis into account given that it affects the probability of the evidence. That, however, is directly at odds with the strictures of methodological falsificationism.

Most of the components of this picture, individually, are familiar. However, what has, arguably, not come clearly into view is the way these pieces hang together. Optimal experimental design, for example, has been around for a long time. It is not new. However, the framing of optimal experimental design typically simply presupposes Bayesianism (as opposed to deriving Bayesian conditionalisation from more fundamental notions as in this paper). This limits its impact in a context where Bayesianism is itself contested by proponents of frequentist statistics. For anyone questioning Bayesianism, the benefits of optimal experimental design formulated in its terms, are ultimately question begging.

In other words, though the parts of the picture presented here are familiar, the key underlying conceptual drivers have remained obscured. Independently of falsificationism, there is a widespread intuition that we should make decisions about what evidence to seek independently of our (current) degree of belief in the hypothesis, and that not doing so is biased and irrational. In fact, the opposite is true by virtue of the standard articulation of rational decision-making. On that standard normative theory, the probability of obtaining the evidence matters. In fact, Popper seemed fundamentally to share that intuition and, share along with it, the intuition that the probability of obtaining the evidence matters to its 'value'. Thus, while contrary in letter, the probabilistic picture is actually congruent in spirit. Importantly, also, these issues are distinct from the choice between frequentist and Bayesian statistics. That confirmation/dis-confirmation is an insufficient basis to drive rational choice about how best to test hypotheses is distinct from the question of how best to analyse the data of whatever experiment we have ultimately decided to pursue. One could (and should) be able to recognise the conceptual flaw inherent in falsificationism regardless of one's personal preference for frequentist versus Bayesian statistics.

References

Blackwell, D., & Dubins, L. (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, *33*(3), 882–886.

Cohen, J. (1994). The earth is round (p<. 05). *American psychologist*, *49*(12), 997.

Crupi, V. (2025). Confirmation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2025). Metaphysics Research Lab, Stanford University.

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2018). Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive science*, *42*(5), 1410–1456.

Fitelson, B. (1999). The plurality of bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of science*, *66*(S3), S362–S378.

Fuchs, R., & Hartmann, S. (2024). Novel predictions for boundedly rational agents: A bayesian analysis. *Proceedings of the annual meeting of the cognitive science society*, *46*.

Gaifman, H., & Snir, M. (1982). Probabilities over rich languages, testing and randomness. *The journal of symbolic logic*, *47*(3), 495–548.

Glymour, C. (2016). Why i am not a bayesian. In *Readings in formal epistemology: Sourcebook* (pp. 131–151). Springer.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Good, I. J. (1966). On the principle of total evidence. *The British Journal for the Philosophy of Science*, *17*(4), 319–321.

Gopnik, A. (1996). The scientist as child. *Philosophy of science*, *63*(4), 485–514.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Hahn, U. (2025). Confirmation bias and the plausible distribution of evidence. *Mind & Society*, 1–18.

Hahn, U., Merdes, C., & von Sydow, M. (2018). How good is your evidence and how would you know? *Topics in Cognitive Science*, *10*(4), 660–678.

Hawthorne, J. (1994). On the nature of bayesian convergence. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1994*(1), 240–249.

Holtz, P., & Monnerjahn, P. (2017). Falsificationism is not just 'potential' falsifiability, but requires 'actual' falsification: Social psychology, critical rationalism, and progress in science. *Journal for the Theory of Social Behaviour*, *47*(3), 348–362. https://doi.org/https://doi.org/10.1111/jtsb.12134

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The bayesian approach*. Open Court Publishing.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, *40*(9), 1098–1101.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *4*(3), 227–241.

Jeffrey, R. C. (1975). Probability and falsification: Critique of the popper program. *Synthese*, *30*(1/2), 95–117.

Karmiloff-Smith, A. (1988). The child is a theoretician, not an inductivist. *Mind & Language*, *3*(3), 183–196.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, *94*(2), 211.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological review*, *96*(4), 674.

Lakatos, I., et al. (1978). The methodology of scientific research programmes. *Philosophical papers*, *1*.

Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, *10*(1), 117094.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in methods and practices in psychological science*, *1*(2), 259–269.

Landes, J., & Williamson, J. (2013). Objective bayesianism and the maximum entropy principle. *Entropy*, *15*(9), 3528–3591.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related bayes' estimates. *University of California Public Statistics*, *1*(11), 277–330.

Levinstein, B. A. (2017). A pragmatist's guide to epistemic utility. *Philosophy of Science*, *84*(4), 613–638.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005.

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

Meder, B., Crupi, V., & Nelson, J. D. (2022). What makes a good query?: Prospects for a comprehensive theory of human information acquisition. In I. Cogliati Dezza, E. Schulz, & C. M. Wu (Eds.), *The drive for knowledge: The science of human information seeking* (pp. 101–123). Cambridge University Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175–220.

Nordgaard, A., & Rasmusson, B. (2012). The likelihood ratio as value of evidence—more than a question of numbers. *Law, probability and Risk*, *11*(4), 303–315.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological review*, *101*(4), 608.

Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford University Press.

Rainforth, T., Foster, A., Ivanova, D. R., & Bickford Smith, F. (2024). Modern bayesian experimental design. *Statistical Science*, *39*(1), 100–114.

Rosenkrantz, R. D. (1992). The justification of induction. *Philosophy of Science*, *59*(4), 527–539.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. oxford university press.

Steele, K., & Stefánsson, H. (2021). Belief revision for growing awareness. *Mind*, *130*(520), 1207–1232.

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In

Bayesian evaluation of informative hypotheses (pp. 181–207). Springer.
Wenmackers, S., & Romeijn, J.-W. (2016). New theory about old evidence. *Synthese*, *193*(4), 1225–1250.
Williamson, J. (2003). Bayesianism and language change. *Journal of Logic, Language and Information*, *12*, 53–97.
Williamson, J. (2010). *In defence of objective bayesianism*. Oxford University Press.