Accelerating scientific progress through Bayesian adversarial collaboration

Andrew W. Corcoran[1,2,*]
andrew.corcoran@monash.edu
0000-0002-0449-4883

Jakob Hohwy[1,2]
jakob.hohwy@monash.edu
0000-0003-3906-3060

Karl J. Friston[3,4]
k.friston@ucl.ac.uk
0000-0001-7984-8909

[1] Monash Centre for Consciousness & Contemplative Studies, Monash University, Melbourne, Victoria, Australia

[2] Cognition & Philosophy Laboratory, School of Philosophical, Historical, and International Studies, Monash University, Melbourne, Victoria, Australia

[3] Wellcome Centre for Neuroimaging, Institute of Neurology, University College London, London, UK

[4] VERSES Research Lab, Los Angeles, California, USA

* Corresponding author

Correspondence
Postal address: Rm 422, Level 4, 29 Ancora Imparo Way, Clayton, VIC 3800, Australia
E-mail address: andrew.corcoran@monash.edu
Twitter handle: @mr_corcorana

Abstract

Adversarial collaboration has been championed as the gold standard for resolving scientific disputes. Although the virtues of adversarial collaboration have been extensively discussed, the approach has gained little traction in neuroscience and allied fields. In this Perspective, we argue that adversarial collaborative research has been stymied by an overly-restrictive concern with the falsification of scientific theories. We advocate instead for a more expansive view that frames adversarial collaboration in terms of Bayesian belief updating, model comparison, and evidence accumulation. This framework broadens the scope of adversarial collaboration to accommodate a wide range of informative (but not necessarily definitive) studies, while affording the requisite formal tools to guide experimental design and data analysis in the adversarial setting. Crucially, this approach enables theoretical models to be scored in terms of a common metric of evidence, thereby furnishing a means of tracking the amount of empirical support garnered by competing theories over time.

Keywords: Adversarial collaboration, Bayesian inference, evidence accumulation, falsification, meta-science, model comparison

> *A frustrated judge in an English (adversarial) court finally asked a barrister after witnesses had produced conflicting accounts, 'Am I never to hear the truth?' 'No, my lord, merely the evidence', replied counsel.*
>
> – Peter Murphy, *A Practical Approach to Evidence*, 1988

1. Introduction

Scientific progress depends on the accumulation of knowledge derived from formal and empirical tests of scientific beliefs [1]. Key to this endeavour is the development of reliable experimental procedures for testing predictions generated under theoretical models (i.e., hypotheses). Typically, such procedures are designed and deployed by scientists interested in garnering evidence for (or against) a particular theory or hypothesis. More rarely, proponents of rival theories band together to formulate experiments capable of adjudicating their disagreements [2–4]. Such *adversarial collaborations* have been championed as the 'gold-standard' for settling scientific disputes, thereby accelerating scientific progress through stringent but fair tests of competing theories [5,6]. However, adversarial collaboration is also attended by various costs and challenges that have hindered its uptake in the cognitive and social sciences [5,7].

This paper aims to address two problems that may have curbed the adoption of adversarial collaboration in neuroscience and psychology – at least until recently.[1] The first problem is conceptual; it concerns what we consider to be a misleading (and overly restrictive) emphasis on the centrality of *falsification* in adversarial collaborative research. Rather than serving to definitively resolve theoretical disagreements in favour of one party or another, we argue that the primary function of adversarial collaboration is to formulate experiments that inform probabilistic (Bayesian) belief updates over a family of theoretical models. This more expansive conception of adversarial collaboration would be particularly salutary for relatively immature fields of inquiry, such as the neuroscience of consciousness.

The second problem is more practical in nature; it concerns the foundational question of how the evidence generated by adversarial experiments should be evaluated and integrated with prior knowledge. We aim to solve this problem by appealing to Bayesian principles of optimal

[1] See, for example, the structured adversarial collaboration projects supported by the Templeton World Charity Foundation's 'Accelerating Research on Consciousness' initiative (https://www.templetonworldcharity.org/accelerating-research-consciousness-our-structured-adversarial-collaboration-projects), and the psychological and social science research supported by the Adversarial Collaboration Project (https://web.sas.upenn.edu/adcollabproject/).

experimental design, evidence accumulation, and model comparison. We show how competing theoretical hypotheses of varying degrees of specificity (*precision*) can be encoded as prior constraints over the parameters of a generative model prescribed by each theory. We then consider how the evidence furnished by inverting such models – using standard variational techniques – can be aggregated across disparate experimental paradigms. Under this integrative framework, adversarial collaboration serves to advance scientific progress one evidence-based belief update at a time.

2. Falsificationism and the *experimentum crucis*

At the heart of many contemporary discussions of adversarial collaboration is a fundamental commitment to *falsificationism* (see, e.g., [5]). In principle, rival theorists should be able to identify certain systematic differences in the empirical observations predicted under their respective theoretical models, and to agree upon suitable methods for obtaining and analysing such data. If all goes well, this analysis will reveal the data to be concordant with one model's predictions, thereby earning a substantial reputational boost for the theory that 'survived' the adversarial challenge unscathed. The defeated party is faced with the unenviable choice of abandoning their 'falsified' model, modifying it in such a way as to accommodate the newly acquired data, or ascribing unfavourable results to previously unforeseen (but not theoretically compromising) factors [7].

This (admittedly simplified) characterisation of adversarial collaboration harkens back to the notion of the *experimentum crucis*, an experiment capable of rendering data that decisively favour one theory or hypothesis over its competitors [8]. Perhaps the most famous example of such an experiment is Eddington's seminal investigation of gravitational lensing (i.e., the deflection of starlight as it passes through the gravitational field of the sun) [9]. Eddington's observations during the 1919 solar eclipse were compatible with Einstein's theory of general relativity [10], which predicted double the degree of light-bending expected under Newton's law of universal gravitation.[2] This result – which catapulted Einstein to international fame [14] –

[2] In fact, the 1919 eclipse results were not so straightforwardly definitive as sometimes portrayed. Although the data collected on Eddington's expedition were concordant with the theory of general relativity, data from one of the telescopes used by another team appeared to challenge it; however, this latter dataset was excluded from analysis on account of technical artefact. This decision would lead to later accusations of bias against Eddington [11] – who was, supposedly, "an enthusiastic proponent of general relativity" (p. 37; [12]) – although recent re-analysis of the data vindicates the original study's conclusions [13]. Aside from this later controversy, it is also noteworthy that publication of the expedition report

greatly influenced Karl Popper's philosophical development of falsificationism [15,16]. Indeed, Eddington's expedition has been highlighted in recent neuroscientific discourse on adversarial collaboration as a paradigmatic example of the sort of experiment that ought to be designed to arbitrate rival theories of consciousness (e.g., [17,18]; see also [19]).

It is notable in light of such discourse that early (and indeed, more recent; e.g., [5,6]) proponents of adversarial collaboration in the psychological sciences generally refrain from lionising the *experimentum crucis* – settling instead for a rather more modest, incremental picture of scientific progress [2,4]. Indeed, adversarial collaboration often fails to reconcile conflicting views [7,20], even in cases where experimental outcomes appear decisive (e.g., [21]). One might argue that the committed theorist's reluctance to concede defeat in the face of disconfirmatory evidence is incidental to the judgement of a broader community of more impartial (or at least, less-personally invested) observers. Nonetheless, scientists' proclivity to "exploit the fount of hindsight wisdom […] when disliked results are obtained" (p. 270) [4] – coupled with the observation that adversarial research tends to generate complex results that neither fully support nor defeat opposing hypotheses [22] – would seem to undercut the notion that adversarial collaboration affords an efficient route to scientific progress via the elimination of deficient theories.

While the *experimentum crucis* might function as a useful ideal when attempting to design studies capable of discriminating amongst competing theoretical hypotheses, emphasis on definitive falsification might prove counterproductive for a number of reasons. First, this attitude may set an unrealistically high bar for many areas of theoretical disagreement, potentially stymying adversarial collaboration at the first hurdle. Indeed, the high-stakes nature of the *experimentum crucis* might go some way to explaining the limited uptake of adversarial collaboration in neuroscience and allied fields over the past two decades. Lowering the stakes such that experiments are designed with a view to furnishing informative – but not necessarily *fatal* – observations might embolden more theorists to collaborate with their adversaries, while making mutually acceptable experimental designs easier to find.

Moreover, the search for an *experimentum crucis* might not be feasible (nor indeed desirable) in certain areas of neuroscientific research. In nascent fields such as the neuroscience of consciousness, for instance, there is widespread controversy about the

---

sparked contemporary scientific debate about alternative theoretical explanations of the lensing effect. Such historical details highlight the inherent difficulty of settling scientific disputes through critical experiments.

fundamental nature of the target phenomenon and how it ought to be investigated [23–26]. While several theories of consciousness have garnered empirical support, recent work indicates that they tend to pick out distinct aspects of consciousness rooted in largely non-overlapping domains: e.g., brain networks, phenomenology, information theory, et cetera [27]. Hence, even if an adversarial experiment were to rule in favour of one theory over another, it may be premature to jettison the defeated candidate altogether: all that has been demonstrated in this situation is a case in which Theory A outperforms Theory B; the converse may be true in another domain. Consciousness research may be better served by adversarial collaborations that evaluate the performance of various theoretical frameworks in multiple domains (e.g., through a more 'holistic' series of experiments designed to test a range of predictions about different sorts of observations; see, e.g., [28]) before arriving at firm abductions about which avenues should be pursued and which excluded from future exploration.

3. Bayesian bets and the value of information

Falsificationism as formulated by Popper characterises the scientific process as the accrual of provisional knowledge in the form of theories that withstand the challenge of empirical *refutation*. The logic of this philosophical doctrine – which continues to find widespread acceptance as a core tenet of modern science amongst cognitive scientists (e.g., [29,30]) – is deeply enmeshed with classical (frequentist) notions of statistical inference. Notably, however, the normative question of how one ought to update one's beliefs following the acquisition of new data is naturally handled by Bayesian inference. Given the often mixed results of adversarial collaborative research (or at least, the mixed interpretations advanced by its protagonists), we propose Bayesian belief updating is more apt for naturalising and conceptualising the role of adversarial collaboration in advancing scientific progress.

We are not the first to conceive of adversarial collaboration in Bayesian terms. Noting the ease with which scientists tend to fall back on auxiliary hypotheses to explain away inconvenient results, Tetlock and Mitchell [31] encouraged adversarial parties to stake 'Bayesian bets' on the outcomes of their experiments (cf. [32,33]). These wagers formalise experimental predictions in terms of a likelihood function defining the probability of various patterns of empirical data under a given hypothesis.[3] The more these likelihoods differ

[3] Technically, this function is a *marginal likelihood*; namely, the likelihood of the data under a given theory or model. This is also known as *model evidence*. The model evidence marginalises over any unknown variables or model parameters; thereby accommodating any sources of irreducible

across competing hypotheses (i.e., the more the likelihood ratio diverges from unity), the greater the potential for empirical data to arbitrate these hypotheses (assuming prior beliefs about the *ex ante* probability of competing hypotheses are not overwhelmingly loaded in favour of one theory).

This Bayesian perspective affords three important insights concerning the nature of adversarial collaboration: First, it naturally accommodates cases in which the data fail to decisively arbitrate between one theory and another (e.g., as a consequence of observing data that were deemed unlikely under both theories). Second, it explains why adversarial parties might arrive at different conclusions concerning the results of their collaboratively designed experiment – proponents of one theory may (informally) assign higher prior probabilities to their favoured hypotheses relative to their adversaries, and may entertain different (posterior) beliefs given the same likelihood function over candidate hypotheses. Without explicit specification of these beliefs, it is impossible to anticipate the magnitude of belief updating that ought to be driven by the empirical data (although prior beliefs can be recovered from posterior beliefs and the empirical likelihood). We postpone further discussion of this issue to Section 5. Third, the relationship between the likelihood ratio and the informativeness of novel observations speaks to a key principle of *optimal experimental design* [34,35], to which we turn next.

One of the most distinctive features of adversarial collaboration is the development of experimental protocols designed to address research questions in a way that satisfies theorists of different stripes. This process has been argued to complement recent advances in the open science movement (e.g., preregistration of predictions, experimental methods, and analysis procedures) by imposing additional constraints on researcher degrees of freedom pertaining to experimental design [5,6]. Participation in adversarial collaboration requires theorists to pursue questions and adopt methods that all parties consider worthwhile, avoiding the temptation to develop studies that 'load the dice' in favour of one's preferred theory [7]. This is to say that the structure of adversarial collaboration is designed to generate empirical results that are informative to the scientific community at large, not merely a subset of scientists working within a particular framework – or limited repertoire of theories.

uncertainty. Under the assumption that all theories are, *a priori*, equally likely, the evidence reduces to the probability of each theory, given the data.

From a normative (Bayesian) perspective, one can construe the task of experimental design as that of inferring the most reliable course of action (policy) for generating informative (epistemically valuable) observations. An informative observation is one that reduces uncertainty about the state of affairs, thereby helping to disambiguate the candidate hypothesis (theoretical model) that provides the best explanation of the generative process of interest. Optimal experimental design thus entails the adoption of procedures that generate data that are expected to reduce uncertainty (*Shannon entropy* [36]) over competing hypotheses (models); or equivalently, to maximise the gain of (Shannon) information about the hidden (latent) causes of observed phenomena. This expected information gain is quantified by the Kullback-Leibler (KL) divergence (i.e., *relative entropy* [37]) between the predictive posterior and prior distributions [35,38,39]. It turns out, mathematically, that the expected information gain is the *mutual information* – between latent (theoretical) causes and empirically observable consequences – expected under a particular experimental design.

What constitutes the most informative experiment within a given context will depend on the family of experiments under consideration. While one theorist may consider a particular experiment the most effective means of disambiguating between predictions generated under two versions of Theory A, another theorist may consider this experiment uninformative as to the broader question of whether Theory A provides a better account of the target phenomenon than Theory B. Adversarial collaboration is valuable (and difficult) to the extent that it overcomes such impasses, providing the conditions for rival theorists to reach consensus on the best experiments for generating the most informative data for all concerned. Notice that this criterion of *informativeness* (i.e., epistemic value) does not depend on whether the data are capable of falsifying one or other theoretical model outright: progress is made to the extent that information is gained about the target phenomenon (i.e., latent causes of data) in the domain shared by all theories at hand.

Seen in this light, adversarial collaboration fosters scientific progress by facilitating the development of highly-informative experiments. Such experiments are characterised by two features: (i) they solicit data about highly uncertain states of affairs (i.e., latent causes about which little is known); (ii) they sample observations that can be reliably mapped to the underlying states that caused them (i.e., states about which something can be learnt). Adversarial collaboration engenders feature (i) by encouraging adversaries to propose experiments that step out of their respective ‘comfort zones’ – that is, to stake risky Bayesian bets about the patterns of data that ought to obtain under novel conditions. This mitigates the temptation to seek evidence consistent with one’s preferred model where it may be easy

to find (e.g., by sampling more certain, less-informative sources; cf. the 'streetlight effect' [40]). Adversarial collaboration also delivers on feature (ii), since joint participation in experimental design should underwrite the selection of methodological procedures expected to promote unambiguous (rather than underdetermined or 'aliased') observations. In other words, prior commitment to a particular experimental policy mitigates the temptation to dismiss inconvenient data as being uninformative.

4. Theory comparison as (Bayesian) model comparison

We have proposed a Bayesian perspective on adversarial collaboration in which the collaborative effort facilitates the selection of appropriate experiments for informing debates about competing theories. On this view, the binary logic of falsificationism is replaced by a more flexible, continuous process of belief updating, in which the probabilities assigned to competing theories depend on the likelihood of observations under their respective hypotheses. However, as alluded to above, Bayesian inference permits adversarial parties to ascribe different prior beliefs about the probability of their favoured theoretical hypotheses; leading to different posterior beliefs – i.e., belief updating – in the face of the evidence (a.k.a., marginal likelihood). While this need not be problematic in itself, it may curb the utility of adversarial collaboration as a mechanism for evincing scientific consensus. In this section, we sketch out a method of theory comparison that aims to formalise the evaluation of theoretical models in the adversarial setting. We then return to the issue of divergent priors.

The approach considered here inherits from Bayesian methods of (hierarchical generative) model comparison [41–43]. Hierarchical Bayesian modelling affords an elegant way to formalise scientific theories as models encoding a hierarchically-structured hypothesis space [44]. In this perspective, higher levels of a model specify more abstract or generic hypotheses, while lower levels generate more concrete or specific hypotheses about the process in question. The hierarchical structure of such models ensures that higher-level predictions impose theoretically-informed constraints on the parameters of lower levels of the hierarchy, thereby influencing the model's predictions about the sorts of data the target process will generate. One can then evaluate candidate models by comparing their capacity to account for empirical observations sampled from the generative process (i.e., experimental data).

To motivate this approach via a toy example, imagine that we performed a psychophysics experiment under two conditions, where each condition included five levels of some experimental factor (e.g., stimulus intensity). Figure 1 illustrates the kind of data one might obtain from such an experiment. In fact, these data are synthetic and were generated by the generative model presented in Figure 2. Having a generative model underneath the data enables us to ask the question: *is there any evidence for a difference in the implicit detection threshold between the two conditions – and, if so, which hypothesis best explains this difference?* Figure 2 illustrates how to assess this evidence, with a special focus on why having a more precise hypothesis can evince greater evidence in its favour.

Practically speaking, the Bayesian approach to theory comparison can be decomposed into three steps: (i) model specification; (ii) model inversion; and (iii) model comparison. First, a generative model must be specified in accordance with the nature of the generative process under examination. For example, if our experimental design entails the collection of two-alternative forced choice (2AFC) responses, a model capable of generating binary choice data could be constructed via the specification of a binomial likelihood function (Figure 2a). This 'generic' model could then be elaborated in various ways according to the particular predictions of alternative theories. This may be achieved by imposing different (theory-specific) priors over critical model parameters; namely, those parameters underwriting distinctive hypotheses concerning the presence (or absence) of experimental effects (Figure 2b). In the case of our 2AFC task, a generic binomial model could be equipped with a prior that permits variability in the patterning of responses (or associated psychometric parameters) across the two experimental conditions (as hypothesised under Theory A), while another model could be equipped with a prior stipulating no difference between conditions (i.e., the *absence* of an experimental effect, as hypothesised under Theory B; see Box 1).

Once models encoding the different hypotheses of competing theories have been specified, they can be fitted to empirical data. Here, we appeal to variational methods based on the Laplace approximation [45–49], as is standard in machine learning and neuroscience (e.g., dynamic causal modelling). These techniques afford an efficient approximation of intractable integrals, rendering an explicit estimate of *model evidence* (a.k.a. marginal likelihood) – an indicator of model quality that outperforms common alternatives such as the Akaike and Bayesian information criteria [50]. Variational methods are widely available through open source software packages such as the Statistical Parametric Mapping (SPM) toolbox (https://www.fil.ion.ucl.ac.uk/spm/).

Bayesian model inversion entails updating prior beliefs to posterior beliefs about model parameters after the model has been exposed to new data (Figure 2c). Bayesian update schemes seek to update prior beliefs by inferring the optimal balance between model *accuracy* (i.e., the discrepancy between predictions and observations) and model *complexity* (i.e., the degrees of freedom required to accommodate observations). This optimal balance is implicit in maximising (log) model evidence via the (negative) variational free energy: a.k.a., the *evidence lower bound* or ELBO [48]. The upshot of this process is a model that fits the observed data as well as possible while remaining as simple (i.e., parsimonious) as possible, thereby mitigating the risk of overfitting parameters to the observations that happened to have been sampled from the generative process during data collection (cf. Ockham's razor [51]).

The Bayesian belief update – realised by model inversion – is scored by the KL divergence between the posterior and prior distributions (i.e., model complexity). Note that, while Bayes-optimal experimental design calls for experiments that are expected to *maximise* the expected KL divergence (and hence precipitate greater information gain), model inversion calls for *minimisation* of the KL divergence, in accordance with the imperative to restrict model complexity (i.e., update the model as little as possible in light of new data). This apparent contradiction is resolved by the intuition that scientists ought to prefer experiments that procure the most informative observations (i.e., resolve the most uncertainty about hidden states), while also preferring explanations that integrate novel observations within the context of prior beliefs as conservatively as possible (i.e., avoid more complicated belief updates than are necessary, given the evidence at hand; cf. the maximum entropy principle [52]).

Having deployed variational methods to invert generative models encoding the predictions of competing theories, one now has access to the free energies that approximate (i.e., bound) the evidence for each model. Model evidence reflects the probability of the observed data given a generative model of those data, and is simply the difference between the accuracy and complexity terms alluded to above. Intuitively, models accumulate more evidence (i.e., their marginal likelihood increases) when their prior predictions accurately characterise empirical observations. In other words, the farther posteriors have to depart from prior beliefs to explain sampled data – i.e., the larger the KL divergence between posterior and prior beliefs – the greater the complexity penalty incurred. Consequently, one can rank models according to their log evidence (as approximated by the variational free energy bound) in order to assess which parameterisation affords the best explanation of the data (Figure 2d).

In terms of scientific progress, this process can be repeated indefinitely, to accumulate evidence — from successive experiments — for plausible theories (see Box 2 for a worked example). This simply entails adding the log evidence for each theory from successive experiments (see Figure 3). Informative experiments ensure that the accumulated evidence for each theory diverges with each new experiment (or not, if all theories provide equally good explanations[4]). There is an accepted semantics for differences in log evidence [41], where a difference of three is usually read as 'strong evidence' for one theory or model relative to another. This is because a log evidence of three corresponds to an evidence ratio of 20:1 (cf. a nominal *p*-value of 0.05 in classical inference).

One might ask whether one can apply the theory comparison procedures described herein to theories that inherit from the *free energy principle* [54]. Indeed, this question was asked by our reviewers. This is a revealing question for two reasons: First, because there is an ongoing adversarial collaboration to precisely do this; namely, to compare variants of predictive processing and active inference with integrated information theory (https://www.templetonworldcharity.org/projects-database/0646). Second, because our Bayesian approach to adversarial collaboration arose as an *application of the free energy principle to theory comparison*. The rationale here is that perception, and indeed active sensing (a.k.a., active inference), follow exactly the same rules and imperatives as evidence-based scientific enquiry [55,56]; namely, soliciting those data that maximise expected information gain [39] – and then evaluating those data through a process of Bayesian belief updating to find the *theory* or explanation that has the greatest evidence. When used to describe sentient behaviour, this process is neatly summarised as *self-evidencing* [57].

5. Circumventing incommensurability

We have seen how Bayesian inference can be harnessed to inform both the design and analysis of adversarial collaborative research projects, focusing in particular on the power of (variational) Bayesian methods to operationalise and evaluate competing theoretical models.

[4] Note that evidence is only meaningful in a relative sense. In other words, one can only compare the evidence for one model in relation to others; e.g., using differences in log evidence. The evidence of any single model has no meaning and can change arbitrarily with, say, the units of measurement of the data. This means there is no 'true' model — there is only the 'best' model from among those models considered. In classical inference, this truism explains why one always compares an alternate hypothesis with a null hypothesis. In Bayesian inference, one can compare an arbitrarily large set of hypotheses or models that may include a null hypothesis. And, interestingly, discover the evidence for the null model is greater than a classically significant alternate model. This is known as 'Lindley's paradox' [53].

In this final section, we explore some more general implications of this framework, foregrounding how it may help to accelerate scientific progress through the comparison of seemingly *incommensurable* theories.

The crucial insight here is that the variational free energy bound on (log) model evidence can be used not only to rank the relative quality of alternative generative models of the same dataset, but also as a general purpose metric that can be aggregated across multiple settings and scales (e.g., sampling units, replication sites, experimental paradigms). In other words, variational free energy affords a common currency for quantifying the evidential support accrued under different theoretical models – currency which can be accumulated over multiple studies and modalities (see Box 2 and Figure 3). This paves the way for a flexible and integrative approach to adversarial collaboration, one that (departing from the ideal of the *experimentum crucis*) permits multiple adversarial experiments involving more or less risky Bayesian bets (e.g., an experiment that tests a highly-constrained set of predictions under Theory A versus a weakly-constrained set under Theory B, and vice-versa. This is illustrated in Box 3, using the worked examples of Box 2.

Notably, the logic of (model) evidence accumulation can also be extended to experiments designed beyond the context of adversarial collaboration. This is particularly valuable when a theory lacks predictions about a given scenario, making it difficult to formulate an experiment capable of rendering informative data from this perspective. This speaks to the concern that progress in nascent research fields – such as the neuroscience of consciousness – may be hampered by the development of multiple theories in the absence of common conceptual frameworks, methodological standards, and explanatory targets [27,29]: namely, theories that may ultimately prove incommensurable with one another. Access to a common measure of evidence affords the opportunity to quantify how well different theories are performing (in terms of how well model predictions are borne out by empirical data), even if such theories were developed and tested "in different worlds" (p. 150) [58], so to speak.

It should be stressed, however, that the scoring of evidence accumulated within independently pursued research programmes may not be sufficient to gauge which theory is making the most scientific progress. This is because independent research generates evidence based on independent datasets, therefore enabling alternative theories to accrue similar quantities of evidence in parallel to one another. While such endeavours can be useful for evaluating and refining theoretical models within the context of a particular research domain or tradition, the unique advantage conferred by the adversarial method – from the perspective of scientific progress – inheres in the discriminative value of comparing

competing theoretical predictions on the *same* empirical observations. Engaging in such activity is guaranteed to be informative in one way or another: either one theoretical model will accrue more evidential support than its competitor(s), or all models will accrue similar amounts of evidence (suggesting that the candidate theories afford equally good or bad explanations of the data, or that their predictions could not be adequately disambiguated under the selected experimental design). In this way, Bayesian adversarial collaboration clarifies which theories are making the most scientific progress, and which theories (or experiments) are most in need of revision.

None of this is to say that the outcome of an adversarial collaboration should be the endpoint of scientific debate. As anticipated by Kahneman (p. 729) [2] – and subsequently borne out by two decades of adversarial collaborative research – adversaries seldom converge on a consensus view upon the completion of their collaboration. Competing interpretations are to be expected irrespective of whether collaboration takes place under a Bayesian framework or not. Theorists are just as likely to draw from the 'fount of hindsight wisdom' when evidence favours a rival theory rather than their own. And indeed, such behaviour is entirely rational from a Bayesian perspective; ascribing surprising results to mitigating factors that explain away the discrepancy between expected and observed results enables theorists to change their minds as little as possible, thereby preserving their prior belief in the fidelity of their model (cf. [59]). This is simply to reiterate the point made in Section 3 that a particular set of empirical observations will not compel the same degree of belief updating amongst all observers, given individual differences in the prior probabilities assigned to competing theories.

Does the propensity to persevere with one's favoured theory in the face of countervailing evidence undermine the supposed benefits of adversarial collaboration? We think not, for two reasons. First, as we have argued, the adversarial setting encourages theorists to stake risky Bayesian bets beyond the epistemic safety of their theoretical 'comfort zone'. Even if theorists are quick to ascribe unfavourable findings to unanticipated factors, unreliable measures, or imprecise predictions made at the margins of their framework (rather than admitting a fundamental deficiency in the framework itself), such outcomes disclose potential avenues for further theoretical and/or methodological development. Valuable information has been gained, and with this information a challenge (and an opportunity for scientific progress) has presented itself.

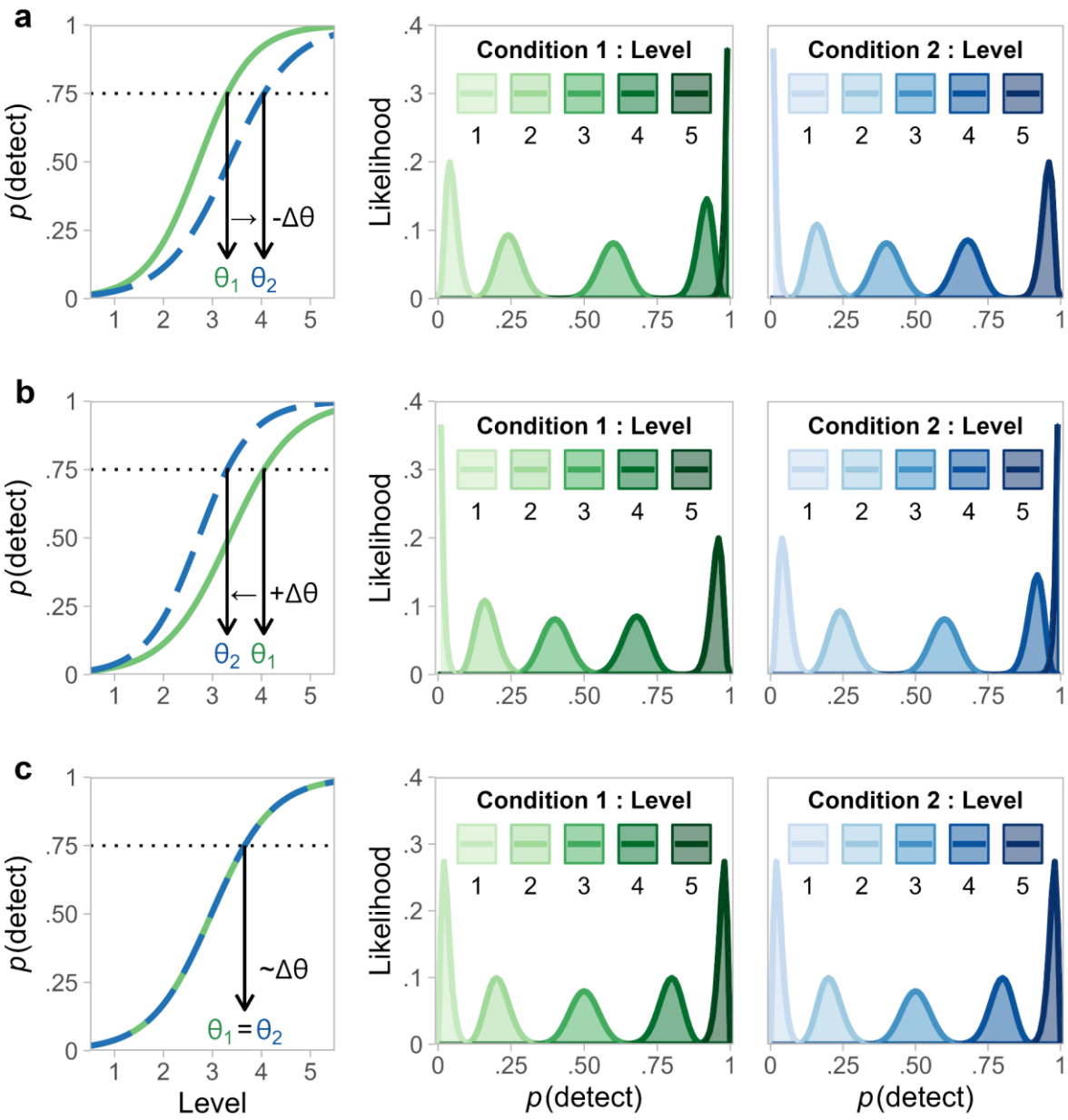
The second reason not to be discouraged by the prospect of lingering disagreement between adversarial parties pertains to the function of adversarial collaboration more

generally. In our view, adversarial collaboration should not be construed as a mechanism for converging on consensus amongst its protagonists. In much the same way as the adversarial legal system is not designed to change the minds of opposing advocates – hinging instead on the arbitration of evidence by an impartial jury of peers – adversarial collaboration seeks to generate the most salient data for the scientific community at large. By fostering rigorous experiments that disambiguate competing hypotheses – pertaining to shared domains of interest – the evidence accrued through Bayesian adversarial collaboration can be exploited to update our beliefs about the merits and prospects of competing theories; ultimately serving to guide individual and collective decisions about resource allocation (e.g., where to invest one's time, energy, and funding). In this way, the knowledge generated through Bayesian adversarial collaboration helps to inform meta-theoretical bets concerning which research policies afford the most promising routes toward scientific progress.

6. Concluding remarks

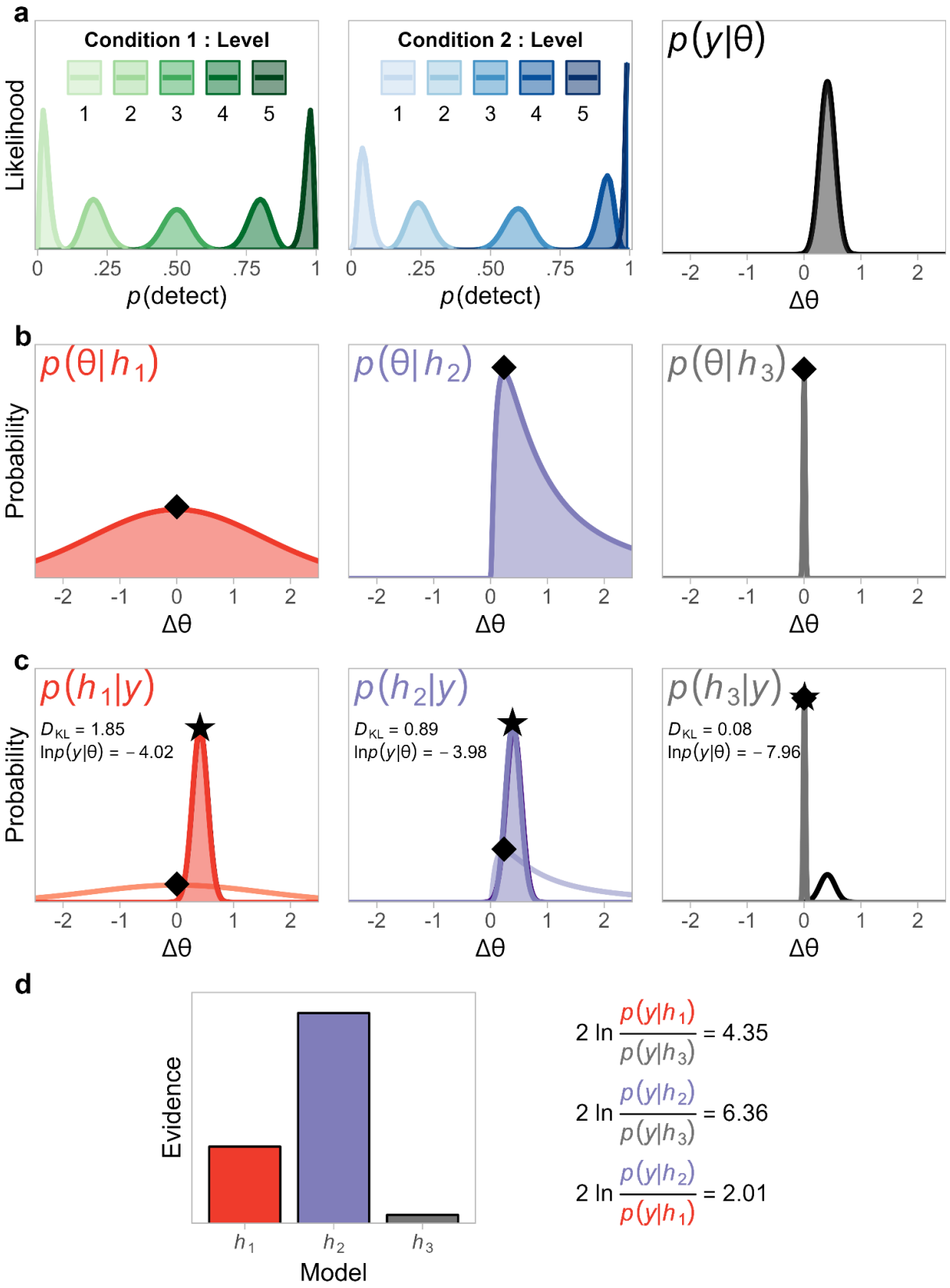
Under ideal conditions, adversarial collaboration helps to advance scientific progress by procuring critical experimental data that convincingly settle theoretical disputes. However, the reality of adversarial collaboration is typically less straightforward: decisive tests of empirical predictions may be difficult to identify; results may fail to definitively arbitrate competing predictions; scientists are wont to interpret data in ways that accord with their own theoretical proclivities. We have proposed that adversarial collaboration should be cast within a (variational) Bayesian framework that facilitates optimal experimental design and the quantification of evidence accrued under theoretical (generative) models. This framework enables evidence to be accumulated over multiple (adversarial and independent) studies, thereby furnishing a simple, common metric of evidential support. As such, it affords a valuable tool for tracking the relative performance of competing theoretical perspectives over time. This information may prove useful for directing resources towards the most promising (i.e., evidence-based) scientific theories, as well as identifying when dominant theoretical frameworks are stagnating and in need of major revision.

Figure 1. Experimental data predicted under three alternative hypotheses.



Left panels exemplify three sets of (synthetic) data that might be obtained from a psychophysics experiment investigating how stimulus detection performance varies across five levels of intensity under two conditions (green and blue functions). The aim of the experiment is to test hypotheses about the effect of some manipulation on the intensity level at which the detection threshold (dotted line) is achieved (where threshold performance at higher intensity levels corresponds to a *reduction* in the sensitivity parameter θ). Middle and right panels depict binomial likelihood functions characterising detection performance at each intensity level in either condition. **(a)** Condition 2 induces a rightward shift in the psychometric function – corresponding to a *decrease* in sensitivity (i.e., negative change in θ) – relative to Condition 1. **(b)** Condition 2 induces a leftward shift in the psychometric function – corresponding to an *increase* in sensitivity (i.e., positive change in θ) – relative to Condition 1. **(c)** The psychometric function remains unchanged across conditions, thus indicating no difference in sensitivity.

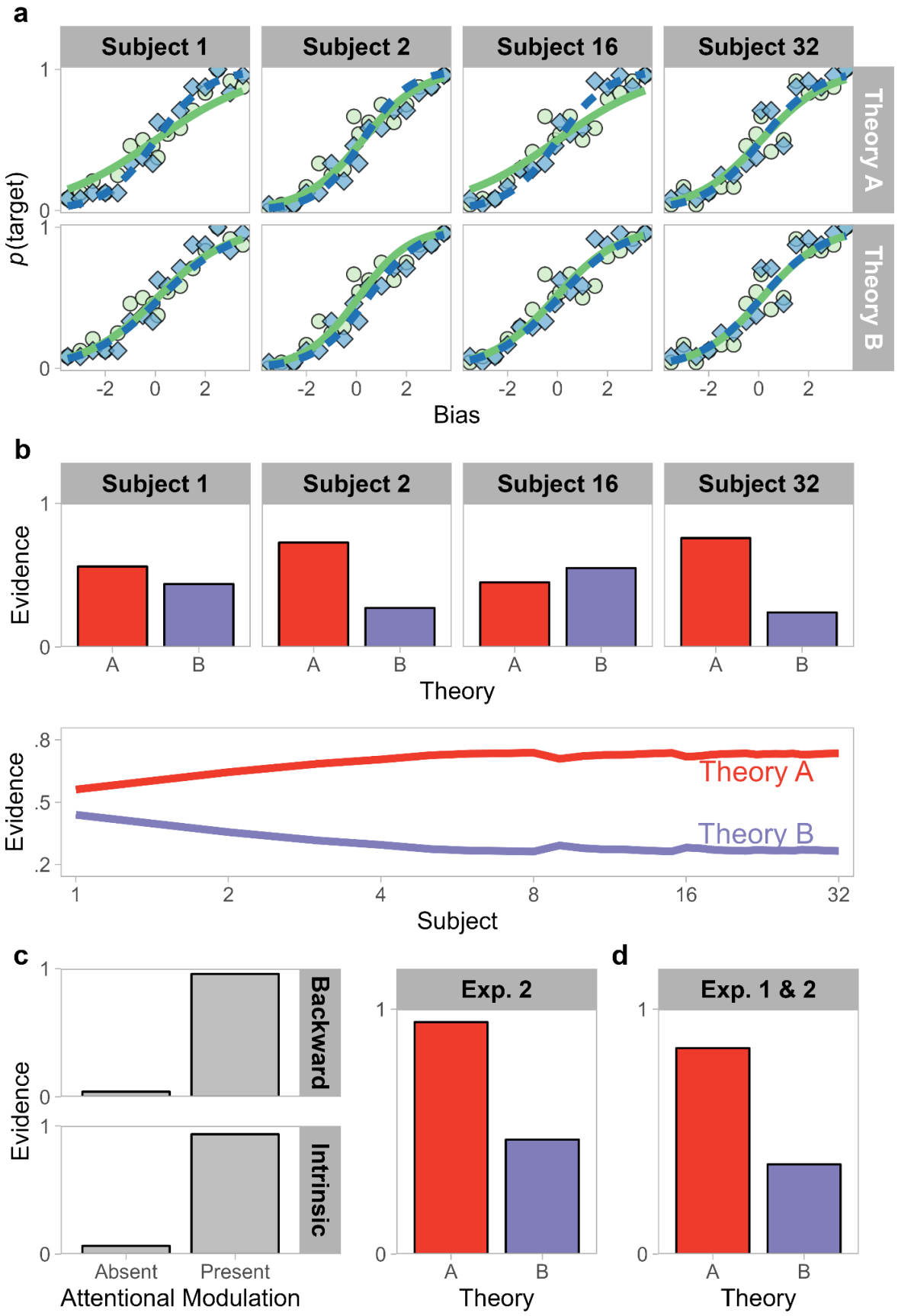
Figure 2. Generative model specification, inversion, and comparison.



(a) Left and middle panels depict binomial likelihood functions describing variation in stimulus detection performance across five levels of stimulus intensity under two

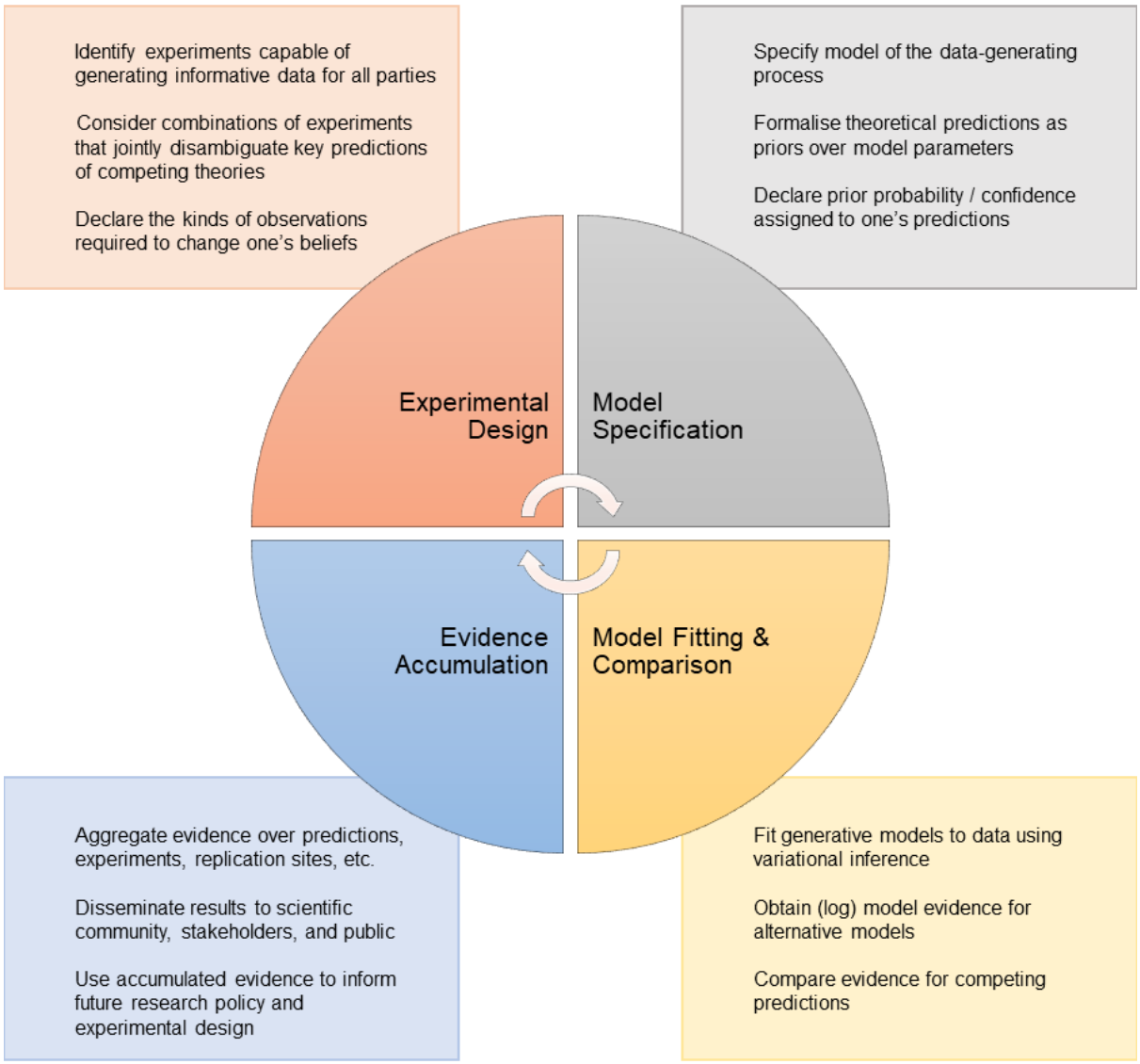
experimental conditions. Right panel depicts the likelihood function $p(y|\theta)$ characterising the change in sensitivity between conditions, $\mathbf{\Delta}\theta$. **(b)** Prior distributions $p(\theta|h_i)$ encode distinct hypotheses about the plausible set of experimental effects expected under competing theories (see Figure 1 for an illustration of how such predictions might manifest in the psychometric and likelihood functions). The location of the modal value of each prior – which signifies the most probable value of $\mathbf{\Delta}\theta$ under each theoretical (generative) model before the data are observed – is denoted by the diamond symbol. The Gaussian prior (left panel) specifies that θ may increase or decrease as a consequence of the experimental manipulation, as reflected by a left- or rightward shift of the psychometric function (i.e., a non-directional hypothesis). The log-Gaussian prior (middle panel) stipulates a more restrictive set of predictions, constraining $\mathbf{\Delta}\theta$ to positive values (i.e., a directional hypothesis). The shrinkage prior (right panel) implies that the manipulation will fail to systematically perturb θ (i.e., a null hypothesis); this model is equivalent to one that pools observations across conditions. **(c)** Posterior distributions $p(h_i|y)$ (filled functions, modal value denoted by the star symbol) obtained via model inversion represent the optimal integration of prior beliefs (unfilled functions) and observed data (black likelihood function). While h_1 and h_2 evince similar degrees of model accuracy (as indicated by their respective log-likelihoods, $\ln p(y|\theta)$), h_2 incurred a lower complexity cost (as scored by the Kullback-Leibler divergence, D_{KL}). By contrast, h_3 remains effectively unchanged by the data. **(d)** Comparison of log marginal likelihoods $p(y|h_i)$ reveals strong evidence of an experimental effect (see [41]). The superiority of h_2 over h_1 highlights the potential advantage conferred by staking riskier ‘Bayesian bets’.

Figure 3. Integration of model evidence over subjects and experiments.



(a) Synthetic data from four (virtual) subjects performing a two-alternative forced choice discrimination task. Green circles represent the proportion of trials in which a target stimulus was selected during the control condition; blue diamonds indicate the proportion of target choices during the experimental condition. Psychometric functions are fit to each dataset according to the prior predictions specified under Theory A (top row) and Theory B (bottom row). While Theory A predicted that the experimental manipulation would promote a change in sensitivity (but not bias), Theory B predicted a change in bias (but not sensitivity). **(b)** Top row displays the (log) model evidence for the generative models fit to the datasets presented in (a); bottom row shows how the relative evidence for either theory evolves as successive subjects are included in the analysis. Here, cumulative estimates of model evidence diverge over the first 8 subjects, fluctuate slightly over the following 8 subjects, and remain stable for the remainder of the sample. This demonstration speaks to the diminishing informational returns availed by further replications of the experiment once uncertainty over hypothesised effects has been adequately resolved. **(c)** Left panel depicts evidence for the presence vs. absence of neurophysiological effects (i.e., the modulation of intrinsic and backwards connectivity) during an attention to visual motion task. While both theories accrue evidence in favour of altered intrinsic cortical excitability, Theory A accumulates additional evidence for correctly positing the modulation of backwards connectivity (right panel). **(d)** Relative evidence for the two theories following integration of the results presented in (b) and (c).

Figure 4. Key stages of Bayesian adversarial collaboration.



This graphic distils the Bayesian approach to adversarial collaboration into four key stages. Although these stages may not be implemented in a strictly sequential fashion (e.g., model specification may reveal insights that motivate revised experimental designs), this cyclical depiction is intended to highlight the iterative, open-ended nature of adversarial collaboration when viewed from the Bayesian perspective.

Box 1. Bayesian theory comparison: A schematic example

Here, we elaborate the example of a psychophysics task in which participants respond to various stimuli presented under two conditions (see Figure 1). Theory A predicts a change in psychometric parameter θ across the two conditions; Theory B predicts no such difference.

First, a generative model encoding the joint probability of the observed data and the hidden states causing them must be defined. This can be accomplished with a binomial likelihood model, specifying the probability of stimulus detection as a function of stimulus kind in either condition (Figure 2a).

Next, this model needs to be equipped with priors that constrain parameter values to be consistent with theoretically-inspired predictions (Figure 2b). The simplest implementation of this contrast is to specify priors that allow θ to vary in the case of Theory A (but not Theory B).

For the model corresponding to Theory A, centring a Gaussian function over θ=0 enables the model to entertain predictions that θ could take a variety of positive or negative values. This set of plausible values could be made more specific by limiting the prior to (e.g.) the positive range of θ. More precise priors (i.e., confining predictions to smaller regions of parameter space) constitute stronger 'Bayesian bets' about a theory's capacity to accommodate novel observations.

For Theory B, placing a highly-precise *shrinkage prior* over θ=0 encodes the prediction that θ does not systematically differ between conditions. This prior effectively 'switches off' θ, rendering a reduced model that can be scored against the Theory A model [60]. The goal of this technique is to evaluate whether the more complex model (Theory A) delivers sufficient improvements in predictive accuracy to justify its additional flexibility. Under variational Bayes, this trade-off is gracefully negotiated via the comparison of (log) model evidence (Figure 2d).

Box 2. Evidence accumulation over effects, subjects, and experiments.

To provide a worked example – that showcases the application of Bayesian theory comparison procedures – we use synthetic data to illustrate evidence accumulation over subjects and modalities. In brief, we simulated two kinds of experiments: a psychophysical experiment and a neurophysiological (fMRI) experiment.

Table 1 summarises the predictions of two *theories* about two *effects* in the two (synthetic) *experiments*. Theorists have the latitude to specify whether an effect was present or not – or to abstain from any prediction. Only by committing to the presence or absence of an effect can evidence be 'claimed' for the theory in question. In other words, making a definitive prediction is a commitment to accepting an increase or decrease in log evidence when comparing generative models with and without the effect in question. In this example, two theorists made specific predictions about both effects in the psychophysics study (a change in *sensitivity* and *bias*, respectively), while the second theorist was noncommittal about one of the effects in the fMRI study (please see below).

Table 1: Theoretical predictions for two effects in two (synthetic) experiments.

	Psychophysical Effects		Neurophysiological Effects	
	Bias	Sensitivity	Intrinsic connectivity	Backward connectivity
Theory A	✗	✓	✓	✓
Theory B	✓	✗	✓	—

Psychophysics study: We generated psychophysical data from 32 (virtual) subjects using a generative model of the sort described in Box 1. These data were generated with condition-specific experimental effects; namely, an increase in the sensitivity (slope) parameter, but no change in bias; i.e., under the predictions of Theory A. Figure 3a shows the (synthetic) data from four subjects, and the corresponding psychometric functions predicted under each theory. The agenda of this example is to illustrate how the evidence for two (or more) theories can be successively accumulated over an increasing number of subjects – showing that an informative picture emerges after a sufficient number of subjects' data are assimilated (Figure 3b).

Neurophysiological study: To illustrate the 'common currency' afforded by (log) model evidence, we next supplemented the results of the psychophysical study with the evidence from a brain imaging study. To do this, we inverted dynamic causal models of fMRI timeseries from a study of attention to visual motion [61]. In this example, the first (correct) theory posited a change in backward connectivity from a higher to a lower region in the visual cortex, in addition to intrinsic (within region) changes in excitability. Conversely, the second theory committed to intrinsic changes in the lower visual area, but declared no prediction for an effect on backward connections. These data were inverted using standard variational procedures [47] – according to open source tutorials on the modelling of these data [62]. Although we only analysed data from one subject, we pretended that this was the result of a group inversion (from 32 subjects) using parametric empirical Bayes [63]. The evidence for models with and without a backwards and intrinsic effect were assessed using an evidence lower bound (a.k.a., variational free energy) in the usual way (Figure 3c). The ensuing evidence for and against the two competing theories was then added to the evidence from the psychophysics study above to give the final results of theory comparison. Unsurprisingly, the evidence was overwhelmingly in favour of Theory A (Figure 3d).

This brief example demonstrates the mechanics of how to transcribe the commitments or predictions of competing theories into an inference procedure that returns an informative, evidence-based assessment of each theory. Key things to note here include the ability of the implicit evidence accumulation to gather evidence over different subjects and data modalities. And, crucially, different kinds of effects that are specified to a greater or lesser degree by each theory (see Box 3); sometimes in an unbalanced fashion. In other words, two (or more) theories may make strong predictions about different kinds of effects and yet these predictions can be used in a complementary way to assimilate evidence for one theory over another. Note further that in Table 1, the commitments of Theory B were less specific than Theory A, insofar as no effects were predicted for changes in backward connectivity. Had the theorist been more specific – and committed to saying that there *were* changes – they would have accumulated more evidence for their theory.

Box 3. The power of precise predictions

One advantage of the Bayesian adversarial collaboration framework is the flexibility it affords to implement various kinds of prediction within a (generally hierarchical) generative model. Theories that propose precise predictions about experimental effects can be specified using bespoke priors that constrain expected parameter values to a narrow range; less-specific predictions (e.g., a positively- or negatively-valenced effect) can be implemented using minimally-informative priors that span the corresponding region of parameter space. Indeed, parameters may even be left entirely unconstrained if a theory lacks any prediction about a particular effect.

Precise predictions constitute riskier Bayesian bets on the outcome of an experiment, since there are many ways in which the model may fail to accurately capture the data. By the same token, data that fit these predictions well furnish compelling evidence of the model's validity. Conversely, weakly-constrained model parameters constitute safer bets, but stand to accrue less evidence irrespective of how the data turn out (since there are many other patterns of data that the model could accommodate equally as well; see Figure 2).

There are a number of ways in which this general scheme can be elaborated. For instance, the confidence each theorist invests in their predictions could be incorporated into the model via the inclusion of hyperpriors over prior constraints on parameters of interest. In this way, the precision accorded to specific predictions could be modulated to reflect stronger or weaker theoretical commitments about putative effects (thus implying stronger or weaker Bayesian bets). One could also incorporate theorists' beliefs about the prior probability of their favoured hypothesis or model; this will constrain the capacity for novel observations to compel belief updates during model inversion. This strategy might be useful for quantifying the magnitude of the 'prediction error' that would be necessary for a theorist to update their beliefs upon observing some data, as compared to an impartial observer (i.e., a scientist who ascribes equal prior probability to each alternative hypothesis).

Acknowledgements

This research was supported by a grant from the Templeton World Charity Foundation (TWCF#0646). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Templeton World Charity Foundation. AWC and JH acknowledge the support of the Three Springs Foundation. KJF is supported by funding for the Wellcome Centre for Human Neuroimaging (Ref: 205103/Z/16/Z), a Canada-UK Artificial Intelligence Initiative (Ref: ES/T01279X/1) and the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3). The authors wish to thank Clement Abbatecola, Melanie Boly, Alex Lepauvre, Andy Mckilliam, Lucia Melloni, Lars Muckli, Niccolò Negro, Umberto Olcese, Cyriel Pennartz, Anil Seth, Giulio Tononi, Peter Zeidman, and three anonymous reviewers for valuable feedback and discussion.

Competing interests

The authors declare no competing interests.

References

1. Bird, A. What is scientific progress? *Nous* **41**, 64–89 (2007).
2. Kahneman, D. Experiences of collaborative research. *Am. Psychol.* **58**, 723–730 (2003).
3. Latham, G. P., Erez, M. & Locke, E. A. Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the Erez–Latham dispute regarding participation in goal setting. *J. Appl. Psychol.* **73**, 753–772 (1988).
4. Mellers, B., Hertwig, R. & Kahneman, D. Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychol. Sci.* **12**, 269–275 (2001).
5. Clark, C. J., Costello, T., Mitchell, G. & Tetlock, P. E. Keep your enemies close: Adversarial collaborations will improve behavioral science. *J. Appl. Res. Mem. Cogn.* **11**, 1–18 (2022).
6. Clark, C. J. & Tetlock, P. E. Adversarial collaboration: The next science reform. in *Political bias in psychology: Nature, scope, and solutions* (Springer, 2022).
7. Cowan, N. *et al.* How do scientific views change? Notes from an extended adversarial collaboration. *Perspect. Psychol. Sci.* **15**, 1011–1025 (2020).
8. Schwartz, D. Experimentum Crucis/Instantia Crucis in the Seventeenth Century. in *Encyclopedia of Early Modern Philosophy and the Sciences* (eds. Jalobeanu, D. & Wolfe, C. T.) 1–5 (Springer International Publishing, 2020). doi:10.1007/978-3-319-20791-9_61-1.
9. Dyson, F. W., Eddington, A. S. & Davidson, C. IX. A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philos. Trans. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character* **220**, 291–333 (1920).
10. Einstein, A. *The collected papers of Albert Einstein. 6: The Berlin years: writings, 1914 - 1917*. (Princeton Univ. Pr, 1996).
11. Earman, J. & Glymour, C. Relativity and eclipses: The British eclipse expeditions of 1919

and their predecessors. *Hist. Stud. Phys. Sci.* **11**, 49–85 (1980).

12. Kennefick, D. Testing relativity from the 1919 eclipse—a question of bias. *Phys. Today* **62**, 37–42 (2009).
13. Gilmore, G. & Tausch-Pebody, G. The 1919 eclipse results that verified general relativity and their later detractors: a story re-told. *Notes Rec. R. Soc. J. Hist. Sci.* **76**, 155–180 (2022).
14. Pais, A. *"Subtle is the Lord... ": The science and the life of Albert Einstein*. (Oxford University Press, 1982).
15. Popper, K. R. *The logic of scientific discovery*. (Routledge, 2010).
16. Popper, K. R. *Conjectures and refutations: The growth of scientific knowledge*. (Routledge, 2002).
17. Melloni, L., Mudrik, L., Pitts, M. & Koch, C. Making the hard problem of consciousness easier. *Science* **372**, 911–912 (2021).
18. Reardon, S. Rival theories face off over brain's source of consciousness. *Science* **366**, 293–293 (2019).
19. Del Pin, S. H., Skóra, Z., Sandberg, K., Overgaard, M. & Wierzchoń, M. Comparing theories of consciousness: Why it matters and how to do it. *Neurosci. Conscious.* **2021**, niab019 (2021).
20. Witkowski, T. Daniel Kahneman: Decision making, adversarial collaboration and hedonic psychology. in *Shaping Psychology* 289–303 (Springer International Publishing, 2020). doi:10.1007/978-3-030-50003-0_15.
21. Matzke, D. *et al.* The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *J. Exp. Psychol. Gen.* **144**, e1–e15 (2015).
22. Cowan, N. The adversarial collaboration within each of us. Comment on Clark et al. *J. Appl. Res. Mem. Cogn.* **11**, 19–22 (2022).
23. Doerig, A., Schurger, A. & Herzog, M. H. Hard criteria for empirical theories of consciousness. *Cogn. Neurosci.* **12**, 41–62 (2021).
24. Francken, J. C. *et al.* An academic survey on theoretical foundations, common

assumptions and the current state of consciousness science. *Neurosci. Conscious.* **2022**, niac011 (2022).

25. Northoff, G. & Lamme, V. Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? *Neurosci. Biobehav. Rev.* **118**, 568–587 (2020).
26. Seth, A. K. & Bayne, T. Theories of consciousness. *Nat. Rev. Neurosci.* **23**, 439–452 (2022).
27. Yaron, I., Melloni, L., Pitts, M. & Mudrik, L. The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nat. Hum. Behav.* **6**, 593–604 (2022).
28. Melloni, L. *et al.* An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLOS ONE* **18**, e0268577 (2023).
29. Cleeremans, A. Theory as adversarial collaboration. *Nat. Hum. Behav.* **6**, 485–486 (2022).
30. Kleiner, J. & Hoel, E. Falsification and consciousness. *Neurosci. Conscious.* **2021**, niab001 (2021).
31. Tetlock, P. E. & Mitchell, G. Implicit bias and accountability systems: What must organizations do to prevent discrimination? *Res. Organ. Behav.* **29**, 3–38 (2009).
32. Hofstee, W. K. B. Methodological decision rules as research policies: A betting reconstruction of empirical research. *Acta Psychol. (Amst.)* **56**, 93–109 (1984).
33. Woodworth, G. G. t for two, or preposterior analysis for two decision makers: Interval estimates for the mean. *Am. Stat.* **30**, 168 (1976).
34. Chaloner, K. & Verdinelli, I. Bayesian experimental design: A review. *Stat. Sci.* **10**, (1995).
35. Rainforth, T., Foster, A., Ivanova, D. R. & Smith, F. B. Modern Bayesian experimental design. (2023) doi:10.48550/ARXIV.2302.14545.
36. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–

423 (1948).

37. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
38. Ginebra, J. On the measure of the information in a statistical experiment. *Bayesian Anal.* **2**, (2007).
39. Lindley, D. V. On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27**, 986–1005 (1956).
40. Demirdjian, D., Taycher, L., Shakhnarovich, G., Grauman, K. & Darrell, T. Avoiding the 'streetlight effect': Tracking by exploring likelihood modes. in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* 357-364 Vol. 1 (IEEE, 2005). doi:10.1109/ICCV.2005.41.
41. Kass, R. E. & Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
42. Kass, R. E. & Steffey, D. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Am. Stat. Assoc.* **84**, 717–726 (1989).
43. Verdinelli, I. & Wasserman, L. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Am. Stat. Assoc.* **90**, 614–618 (1995).
44. Henderson, L., Goodman, N. D., Tenenbaum, J. B. & Woodward, J. F. The structure and dynamics of scientific theories: A hierarchical Bayesian perspective. *Philos. Sci.* **77**, 172–200 (2010).
45. Daunizeau, J. The variational Laplace approach to approximate Bayesian inference. Preprint at http://arxiv.org/abs/1703.02089 (2018).
46. Fox, C. W. & Roberts, S. J. A tutorial on variational Bayesian inference. *Artif. Intell. Rev.* **38**, 85–95 (2012).
47. Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J. & Penny, W. Variational free energy and the Laplace approximation. *NeuroImage* **34**, 220–234 (2007).
48. Winn, J. M. & Bishop, C. M. Variational message passing. *J. Mach. Learn. Res.* **6**, 661–694 (2005).

49. Zeidman, P., Friston, K. & Parr, T. A primer on Variational Laplace (VL). *NeuroImage* 120310 (2023) doi:10.1016/j.neuroimage.2023.120310.
50. Penny, W. D. Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage* **59**, 319–330 (2012).
51. Jefferys, W. H. & Berger, J. O. Ockham's razor and Bayesian analysis. *Am. Sci.* **80**, 64–72 (1992).
52. Jaynes, E. T. Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
53. Lindley, D. V. A statistical paradox. *Biometrika* **44**, 187–192 (1957).
54. Friston, K. J. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
55. Parr, T. Inferring what to do (and what not to). *Entropy* **22**, 536 (2020).
56. Gregory, R. L. Perceptions as hypotheses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **290**, 181–197 (1980).
57. Hohwy, J. The self-evidencing brain. *Noûs* **50**, 259—285 (2016).
58. Kuhn, T. S. *The structure of scientific revolutions*. (University of Chicago Press, 1996).
59. Yon, D., de Lange, F. P. & Press, C. The predictive brain as a stubborn scientist. *Trends Cogn. Sci.* **23**, 6–8 (2019).
60. Friston, K., Parr, T. & Zeidman, P. Bayesian model reduction. (2018) doi:10.48550/ARXIV.1805.07092.
61. Büchel, C. & Friston, K. J. Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cereb. Cortex* **7**, 768–778 (1997).
62. Zeidman, P. *et al.* A guide to group effective connectivity analysis, part 1: First level analysis with DCM for fMRI. *NeuroImage* **200**, 174–190 (2019).
63. Zeidman, P. *et al.* A guide to group effective connectivity analysis, part 2: Second level analysis with PEB. *NeuroImage* **200**, 12–25 (2019).