European Journal for Philosophy of Science (2023) 13:28
https://doi.org/10.1007/s13194-023-00533-w

PAPER IN THE PHILOSOPHY OF THE SCIENCES OF MIND AND BRAIN

Critique of pure Bayesian cognitive science: A view from the philosophy of science

Vincenzo Crupi[1] · Fabrizio Calzavarini[1]

Received: 10 January 2023 / Accepted: 28 May 2023

Abstract
Bayesian approaches to human cognition have been extensively advocated in the last decades, but sharp objections have been raised too within cognitive science. In this paper, we outline a diagnosis of what has gone wrong with the prevalent strand of Bayesian cognitive science (here labelled pure Bayesian cognitive science), relying on selected illustrations from the psychology of reasoning and tools from the philosophy of science. Bayesians' reliance on so-called method of rational analysis is a key point of our discussion. We tentatively conclude on a constructive note, though: an appropriately modified variant of Bayesian cognitive science can still be coherently pursued, as some scholars have noted.

Keywords Bayesian cognitive science · Rational analysis · Is-ought · Predictivism

1 Introduction

Bayesian approaches to human cognition have been extensively advocated in the last decades, but sharp objections have been raised too within cognitive science (e.g., Baron, 1991; Bowers & Davis, 2012a, 2012b; Eberhardt & Danks, 2011; Elqayam & Evans, 2011; Glymour, 2007; Marcus & Davis, 2013, 2015; Sloman & Fernbach, 2008). In this paper, we outline a diagnosis of what has gone wrong with a prevalent strand of Bayesian cognitive science (here labelled *pure* Bayesian cognitive science), relying on selected illustrations from the psychology of reasoning and tools from the philosophy of science. Bayesians' reliance on so-called method of rational

Vincenzo Crupi
vincenzo.crupi@unito.it

Fabrizio Calzavarini
fabrizio.calzavarini@unito.it

[1] Department of Philosophy and Educational Science, University of Turin, Via Sant'Ottavio 20, 10124 Turin, Italy

Published online: 20 June 2023

analysis is a key point of our discussion. We tentatively conclude on a constructive note, though: an appropriately modified variant of Bayesian cognitive science can still be coherently pursued, as some scholars have noted.

2 First exhibit: The Wason task and information search

Consider the (abstract) Wason task, a widely known puzzle from the experimental study of human reasoning (Wason, 1966). Participants are presented with four cards, showing E, K, 2, and 7, respectively. They are told that each card has a letter on one side and a number on the other. Participants are asked to say which cards they would turn over in order to find out whether the following conditional statement is true: (h) *if a card has a vowel on one side, then it has an even number on the other side*.

As it turns out, almost all subjects select the first card (E), a majority also selects the third card (2), only a few select the fourth one (7), and almost nobody selects the second one (K). According to Wason (1966), as is well-known, these results clearly indicated biased reasoning. This is because most participants tend to select a card that is apparently useless for discovering the truth or falsity of the conditional statement h (the third card), but most participants also fail to select a card that is crucial to that effect and equally accessible (the fourth). Turning the first and the fourth card is considered useful in this analysis because, logically, these cards can potentially falsify the hypothesis at issue (by possibly revealing an even number and a vowel, respectively), whereas the other two cards cannot provide any refuting evidence for that hypothesis. In this view, the prevalent pattern of responses suggests that people have a biased tendency to look for "positive instances" of a hypothesis (a form of confirmation bias), thus to look for a vowel behind 2 (third card), and fail to carry out or appreciate the logically valid inference form *modus tollens*, thus missing the opportunity to check whether there's a vowel behind 7 (fourth card). (See Ragni et al., 2018 for a recent and comprehensive discussion.)

If you are a Bayesian, however, you may wonder if there's a better interpretation of this phenomenon (Oaksford & Chater, 1994, 2003, 2007; Nickerson, 1996; Fitelson & Hawthorne, 2010). Maybe some latent subtlety is implied when people face the abstract Wason task. You notice that the four card problem does not look like a matter of simple logic. It requires a model of the rational search for evidence, such as the expected reduction of uncertainty (viz. entropy; see Crupi et al., 2018b) or perhaps the expected reduction of epistemic inaccuracy (as measured by a strictly proper scoring rule; see Vindrola & Crupi, 2021). You postulate that relevant probabilities (say, the probability of finding an even number behind the letter E) can be derived as if the four cards were randomly and independently drawn from a larger deck. Is the conditional statement h true *as concerns such larger deck* (the whole population of cards, as it were)? And which one of the four sampled cards, if turned over, would provide the most useful evidence to find out *that*? This seems a meaningful interpretation of the Wason task, for a Bayesian agent. As it happens, the observed rank of choice propensities (E > 2 > 7 > K) is fully recovered by the expected reduction of either uncertainty or inaccuracy, as long as vowels and even numbers are assumed to be *rare* in the background population (larger deck), a typical illustration being $P(vowel) = 0.22$ and $P(even) = 0.27$ (see Oaksford & Chater, 2003, p. 298). You may reach the conclusion that "subjects behave as Bayesians with the rarity assumption" (Oaksford & Chater, 1994, p. 627).

3 Pure Bayesian cognitive science

We submit our first exhibit above as representing a cornerstone example of contemporary Bayesian cognitive science (BCS) in its most prominent form (Chater et al., 2010; Griffiths et al., 2008; Oaksford & Chater, 2007). For the purposes of the present contribution, we will have to characterize this research program more precisely, and give it a more specific label. We will thus call it *pure* Bayesian cognitive science (PBCS), and submit that the following claims hold for PBCS.

(1) It draws on the Bayesian framework and formal machinery to account for cognitive phenomena documented by behavioral data.
(2) It regards Bayesianism as a compelling general model of rational inference and judgment under uncertainty.
(3) It is committed to a remarkably weak notion of evidential support from the data.
(4) It is bound to imply a subtle but important violation of so-called *is / ought* divide.[1]
(5) It regards humans as essentially rational.

(1) and (2) should not be contentious, we assume. From the work of leading figures of PBCS such as Chater, Oaksford, Tenenbaum, and Griffiths, (1) is evident and (2) is also quite explicit. Concerning (5), note for instance that according to Oaksford and Chater (2007, p. 19) "to show how empirical data on human reasoning can be reconciled with the notion that people are rational" is "the central goal" of their work. The daunting theoretical latitude of the notion of rationality becomes tractable here, at least for our purposes, again due to (2): to count as rational, human behavior should comply appropriately with the benchmark of probabilistic principles of inference and judgment. We expect points (3) and (4) to look much more puzzling in comparison with the others. As they turn out to be important for our argument, we will now explain better their content and why we take it to be tenable that they in fact apply to PBCS. This will require a discussion of "the method of rational analysis".

[1] As we will see, the relevant violation amounts to regarding normative considerations as relevant to decide a descriptive issue, namely, the status of a cognitive model of actual behavior (see Elqayam & Evans, 2011 for a similar discussion). Such inferential move from *ought* to *is* has opposite direction as compared with the traditional *is / ought* fallacy, in which normative conclusions are drawn from purely descriptive considerations.

4 The vagaries of rational analysis

The textbook reference for "rational analysis" is Anderson (1990, 1991). Rational analysis (RA from now on) of a cognitive phenomenon is normally presented as a (iterative) procedure in a series of steps (see. e.g., Anderson, 1990, p. 473; Anderson, 1991, p. 29), along the following lines.

(i) Precisely specify what are the goals of the cognitive system.
(ii) Develop a formal model of the environment to which the system is (taken to be) adapted.
(iii) Make the minimal assumptions about computational limitations.
(iv) Derive the optimal behavioral function given items (i) to (iii).
(v) Examine the empirical literature to see if the predictions of the behavioral function are confirmed.
(vi) If the predictions are off, iterate, revising and refining the theory.

As a matter of fact, the connection between PBCS and RA has been consistent. RA has been often explicitly invoked (e.g., Dubey & Griffiths, 2020; Goodman et al., 2008; Oaksford & Chater, 2007) and — one could argue —it has been largely presupposed and employed as an appropriate research strategy in the PBCS literature. To illustrate, our understanding of how the Bayesian approach to the abstract Wason task instantiates the rational analysis method is roughly as follows (also see Oaksford & Chater, 1994, pp. 625–626). As concerns (i), the cognitive system is assumed to aim at minimizing uncertainty (alternatively, epistemic inaccuracy) in inquiry. As for (ii), the relevant environment is represented by a sampling model (of the four cards from a larger population). This allows for the derivation of the optimal behavioral function (iv), in fact even without substantial additional computational constraints (iii) — *provided*, however, that some key model parameters are fixed. In fact, consideration of basic empirical findings (v) (the ranking of choice propensities for the four cards) substantially constrains parameter setting, thus leading to the rarity assumption as a specific refinement of the general Bayesian analysis (vi) (see Vindrola & Crupi, 2021 for more details).

But why should anyone follow rational analysis to understand human cognition? According to a recent and important reconstruction (Icard, 2018), the "original motivation for rational analysis" is "to help guide the search for cognitive models" when faced with "a problem of *identifiability*" (pp. 2–3):

> The problem is two-fold. First, there is an obvious question of where to begin the search for high-level cognitive models. Second, there are often cases in which competing models cannot be distinguished by available measurement.

We take this passage to be revelatory. It implies that the method of RA is supposed to address *two* issues at once. And in fact, the "problem of identifiability" is best seen as conflating two putative problems. One has to do with theory *construction*, the other one with theory *assessment*. Accordingly, sorting out this distinction generates two distinct interpretations of rational analysis. For reasons to be clarified shortly, we will call them *constructive* and *inferential*, respectively.

5 Two sides of rational analysis

The *constructive* interpretation of RA is, we submit, the following. You *start out* with the working hypothesis (however motivated) that human cognition complies with Bayesian rationality. You then select a certain class of cognitive tasks as your research target (say, information search, memory retrieval, spatial reasoning, …), aiming at an account of some established phenomena in that domain. Given the breadth and generality of the Bayesian framework, a diligent and intelligent application of the prescriptions of rational analysis will essentially *guarantee* that you'll end up with a workable model and that the known phenomena will be recovered as implications of your model. In essence, this is nothing but a specific consequence of Duhemian underdetermination in scientific theorizing (see Laudan, 1990; Crupi, 2020). Indeed, a plurality of outcomes will typically be possible through the process, depending on several modelling choices made along the way. But in any event, *this* use of RA plays a key "heuristic" role in a Lakatosian sense (see Lakatos, 1978): for someone whose premises embed tenets (1), (2) and (5) of PBCS (see above), RA does lead to a solution of the theory-construction problem. No presentation of RA relates it to Lakatos to the best of our knowledge, but if a Lakatosian philosopher of science had to try a characterization of the "positive heuristic" of PBCS as a research program (much as Lakatos did with Newtonian physics, for instance), well, a very close approximation to RA would ensue, or so we suggest (also see Baron, 1991 on this point).[2]

What we take to be the *inferential* interpretation of RA yields a rather different story. Here is a sketch of how it operates. Much as before, you target certain phenomena, C, in a given class of cognitive tasks. Typically, there will exist some non-Bayesian account of such phenomena, call it A. Following the steps of RA, you develop a Bayesian account, B, of the same phenomena. You then consider that "competing models", A and B, "cannot be distinguished by available measurement" (Icard, 2018, p. 3), because by hypothesis now both A and B can account for C. Note that, due to how RA proceeds, the Bayesian account of the data will typically *be driven by the data themselves* through tailored specifications of model and parameters (recall the high level of generality of the Bayesian framework and the crucially "iterative" nature of RA!). As a consequence, the claim of "empirical underdetermination" across competing models (here, A and B) relies, if implicitly, on a collapse of empirical support on plain accommodation of the data. Adopting a terminology from Worrall (2011), "empirical equivalence" is thus reduced to "data equivalence", conveying a suspiciously weak idea of evidential support which is squarely at odds, in particular, with various strands of a "predictivist" view of scientific confirmation (see Crupi, 2020). The collapse of support on accommodation is unjustified because

[2] According to Lakatos, as is well known, positive heuristic consists in a set of rules that helps one in the application of the research program and in the formation of fallible extensions of its "hard core"—that is, in the creation of the protective belt of auxiliary hypotheses: «the positive heuristic consists of a partially articulated set of suggestions or hints on how to change, develop the "refutable variants" of the research programme, how to modify, sophisticate, the "refutable" protective belt» (Lakatos, 1978, p. 50).

mere data equivalence (fit with the data by each of two competing models or theories) is known to have very limited epistemological import in itself. In particular, if one model predicts a phenomenon that another model only accommodates, the former but not the latter will be supported, data equivalence notwithstanding.[3]

Once the "empirical indistinguishability" claim is put forward, however, a case can be made that *B* is still better than *A* for theoretical reasons, to wit, because it reconciles human behavior with compelling standards of rationality. Sometimes this move is accompanied by the claim that the proposed Bayesian model has slightly better empirical fit or offers a more unified account as compared to the alternatives, but the emphasis is clearly put on its rational grounding. Here is a good example of this interpretative pattern as regards the analysis of inductive generalization:

> Our Bayesian model offers a modest but consistent quantitative advantage over the best similarity-based models of generalization, and also predicts qualitative effects of varying sample size that contradict alternative approaches. *More importantly*, our Bayesian approach has a principled rational foundation [...]. In contrast, the similarity-based approach requires arbitrary assumptions [...] that have no apriori justification (Sanjana & Tenenbaum, 2003, p. 65, emphasis added).

Very similar appeals to the rational foundation of Bayesian models in a comparative assessment can be found elsewhere (e.g., Heit, 1998; Kemp & Tenenbaum, 2003; Kemp et al., 2007; Tenenbaum & Griffiths, 2001). So, in inferential RA, one goes on pointing out that *B* is to be favored over *A* because *B*, unlike *A*, is an empirically adequate descriptive model that is *also* supported by a powerful and general normative justification.[4] In this way, the normative status of a theory (an "ought" issue) is meant to contribute to the assessment of the theory as a candidate high-level description of behavior (an "is" issue).

6 Cutting the knot

The constructive interpretation and use of RA is entirely legitimate. It serves the purposes of someone who (for whatever reason) starts out with assumptions which uncontroversially characterize PBCS — (1), (2), and (5) above — and pursues

[3] The details of the distinction between prediction and accommodation are themselves a matter of discussion and may vary in subtle ways: see Barnes (2022) for a valuable survey, and Crupi (2023) for a specific proposal and application to a major historical case.

[4] Some advocates of Bayesian cognitive science explicitly appeal to the normative status of Bayesian inference as supported in usual ways, such as Dutch book arguments (see, e.g., Hahn, 2014, pp. 6–10). Other scholars distinguish "normative rationality", which concerns general standards of correct performance, from "adaptive rationality", which is «defined with respect to a specific environment» (Anderson, 1990, p. 35; but see Oaksford & Chater, 2009 for an explicit attempt to collapse the two notions). In this second sense, specific Bayesian models are rationally/normatively justified based on their optimality with respect to the computational goals of the agents (e.g., Griffiths & Tenenbaum, 2009) or based on their optimal trade-off between decision accuracy and cognitive costs and limitations (e.g., Lieder & Griffiths, 2020).

the development of such research program to accommodate more and more of the empirical basis of the study of human cognition. Such variant of RA is also inferentially *inconsequential*, because the methodological assessment of whether and to what extent such development provides significant support to the overarching program against competing views remains *out of the scope* of this theory-construction process.

The inferential interpretation and use of RA, as we sketched it above, is much more intricate and contentious. It is meant to deliver an *argument* in favor of the rationality of humans in a Bayesian perspective. The starting point is indeed the construction of a Bayesian account of the target phenomena (say, the Wason selection task), to be compared with alternative perspectives (say, Wason's original interpretation). Then the strategy relies on a weak notion of empirical support (point 3 above) to motivate the appeal to considerations beyond "available measurement". The crucial consideration in favor of the Bayesian analysis (and thus of the rationality of behavior, point 5) turns out to be its distinctive normative status (point 2), so that an "ought" statement is taken to contribute (defeasible) support to a claim in the descriptive study of behavior and cognition (point 4). In a slogan, the Bayesian rationality of humans is a *premise* in constructive RA, while it is meant to be a *conclusion* in inferential RA.

We have now explained why we believe that the inferential interpretation and use of RA involves a weak notion of evidential support from the data and a violation of so-called is / ought divide. Our conclusion that points (3) and (4) above apply to PBCS relies on the additional assumption that work in the PBCS program does often embed RA in the inferential sense. Admittedly, endorsement of the inferential variant of RA is seldom explicit and transparent.[5] However, the literature suggests that the constructive and inferential variants of RA are recurrently conflated in PBCS, and indeed the inferential aspect of RA is sometimes rather overtly advocated, like in this important example (Hahn, 2014, p. 10):

> [T]he point of the approach is a methodological one: rational models aide the disambiguation between competing theories [...]. [T]his gives such models [...] a special status, above and beyond degrees of 'model-fit' and so on.

To be sure, the normative status is not the *only* theoretical consideration that can be invoked to solve the comparative assessment in inferential RA. As a matter of fact, the unifying power of the Bayesian models in accommodating various sets of empirical data (see Colombo & Hartmann, 2017 for a critical discussion), as well as their adaptive optimality with respect to evolutionary landscapes (see Bowers &

[5] In this sense, one might regard inferential RA as an idealization that may not be implemented exactly by any particular study, much as it happens with the idea of "imperial" (vs "local") Bayesianism in Mandelbaum (2019). The critical point, however, is that this notion captures a tendency which is actually present in the Bayesian cognitive science literature – indeed, several analyses closely resemble inferential RA in its pure form (e.g., Oaksford & Chater, 1994; Heit, 1998; Tenenbaum & Griffiths, 2001; Sanjana & Tenenbaum, 2003; Kemp & Tenenbaum, 2003; Kemp et al., 2007). At the same time, our reconstruction offers a coherent philosophical basis for some sparse objections raised by critics of the Bayesian approach, as we will see in Section 7.

Davis, 2012a for a critical discussion) have often been often provided as reasons to prefer Bayesian over non-Bayesian accounts. This is particularly true for sub-fields of Bayesian cognitive science such as perceptual (see Rescorla, 2015) and sensorimotor psychology (see Rescorla, 2016), where comparative evaluations with respect to rationality standards are less relevant.

In the empirical study of reasoning and decision-making, however, both in the Bayesian and non-Bayesian tradition, normative considerations have proven to be pervasive and to have generated intense debates (see Elqayam & Evans, 2011; Baron, 2012; Elqayam & Over, 2016; Achourioti et al., 2014; Crupi & Girotto, 2014; Hahn, 2014; Oaksford, 2014). It has been argued that normative standards operate implicitly at many levels in the standard practice of the psychology of reasoning – for example, by constraining the choice of the a priori assumptions at the computational level of analysis (*prior rule bias*), or by influencing the way in which findings are reported and interpreted on the processing level (*interpretation bias*) (Elqayam & Evans, 2011; also see Gigerenzer, 1991). It has also been argued that the implicit appeal to normative standards, especially Bayesian norms, is an ineliminable component of our interpretative practices (Oaksford, 2014), making the departure from the *is* / *ought* divide a constitutive feature of the empirical evaluation of reasoning and decision making. Moreover, the appeal to (Bayesian) optimality/rationality has been invoked as an integral part of explanations at the computational level in David Marr's terms, where theorists must specify not only *what* function is computed by the cognitive system but also "why" the cognitive system operates the way it does (e.g., Griffiths et al., 2012a).

Appeals to Marr's methodology are particularly pervasive in PBCS and they deeply affect how empirical investigations are framed and conducted. Here the underlying assumption is that only rational models are suited to provide adequate explanations at the computational level, in that they specify why the supposedly computed function is normatively appropriate for the task at hand. In this sense, "rational model" and "computational model" tend to be used synonymously (see, e.g., Griffiths & Tenenbaum, 2009, pp. 665–666; Griffiths et al., 2012a, pp. 263–264).[6] In turn, computational analyses are supposed to constrain the lower levels of explanation, where specific hypotheses about cognitive and neural processes are developed («[w]hatever form those cognitive and neural processes take, they need to approximate the solution to the computational problem», Griffiths et al., 2012a, p. 264). As a consequence, Bayesian cognitive hypotheses turn out to be systematically preferred over non-rational alternatives for they allegedly comply with Marr's framework in a distinctive way (e.g., Kemp & Tenenbaum, 2003; Kemp et al., 2007) and «provide a computational-level description *and justification* of why some phenomena occur» (Heit, 1998, p. 271, emphasis added). For the same

[6] For instance, in Griffiths, Vul, and Sanborn's reading (2012a) of Marr, the computational level specifies «the ideal solution to an abstract statistical problem that people must solve: Given the decision that must be made, how *should* people use the limited available information?» (p. 263). The appeal to rationality/optimality is meant to address the "why" component of explanations at this level: «that the solutions are optimal licenses a particular kind of explanation […] allowing us to assert that the match between the solution and human behavior may be why people act the way they do» (Griffiths et al., 2012a, p. 415).

reason, the comparative assessment is sometimes explicitly restricted to predictions generated by psychological models that have a rational foundation (e.g., Griffiths & Tenenbaum, 2009) or to cognitive algorithms that approximate ideal Bayesian solutions (e.g., Griffiths et al., 2012a).[7]

None of the considerations above appears to be conclusive as an epistemic justification of inferential RA anyway. For example, the appeal to the principle of charity has been severely criticized as a philosophical reason to systematically prefer normative (e.g., Bayesian) over non-normative accounts of cognitive phenomena (see Stein, 1996, Ch. 4). Similarly, it has been claimed that there is no principled reason why explanations at the computational level in Marr's terms should contain an ineliminable appeal to rationality/optimality (e.g., Elqayam & Evans, 2011). Quite on the contrary, it seems plausible to insist that «competence-level explanations [are] descriptive, "is"-type theories, rather than normative, "ought"-type theories» (*ivi*, p. 239).

7 A philosophical reconstruction

A more comprehensive survey of the debates mentioned above will have to wait for another occasion. For the moment, we will claim indirect support for our diagnosis, pointing out that it contributes a coherent philosophy of science reconstruction for important but sparse complaints raised by others (see Baron, 1991; Bowers & Davis, 2012a, b; Eberhardt & Danks, 2011; Elqayam & Evans, 2011; Glymour, 2007; Jones & Love, 2011; Marcus & Davis, 2013; Sloman & Fernbach, 2008; Tauber et al., 2017).[8]

For instance, several scholars have complained that strong, unwarranted appeals to normativity and/or optimality are frequently invoked in the Bayesian psychological literature (e.g., Bowers & Davis, 2012a, b; Marcus & Davis, 2013) whereas such considerations are simply irrelevant within the descriptive framework of cognitive science (Tauber et al., 2017). Bayesian models have also been criticized for being too unconstrained; because «there are too many arbitrary ways that priors, likelihoods, utility functions, etc., can be altered in a Bayesian theory post hoc», as observed by Bowers and Davis (2012a, p. 394), «these models […] account for almost any pattern of results», resulting is nothing more than unfalsifiable "just-so-stories". Moreover, it has been argued that Bayesian models rarely perform better than alternative, non-Bayesian models in terms of predicting human performance in reasoning tasks, and they are often preferred over alternatives just for a sort of strong confirmation bias (Bowers & Davis, 2012a).

[7] Griffiths and Tenenbaum's analysis of causal induction offers a clear example of this methodological stance: «[…] our aim to provide a computational level account of causal induction […] influences the kind of models that we use for comparison. In this article, our emphasis is on comparison of the predictions of our accounts to those of other rational models» (Griffiths & Tenenbaum, 2009, p. 666).

[8] We focus here on complaints (and responses) that remain at the computational level in David Marr's terms, which are most relevant in the present context. We shall not discuss other interesting topics such as the critique that Bayesian models does not engage with algorithmic or mechanistic explanations (e.g., Bowers & Davis, 2012a; Jones & Loves, 2011), or the debate about realist *vs* instrumentalist interpretations of Bayesian models (e.g., Colombo & Seriès, 2012; Zednik & Jäkel, 2016; Colombo et al., 2021).

Note that none of the criticisms just mentioned is decisive in itself, as several scholars have noted. For instance, normative considerations might be indeed relevant in descriptive psychology if the aim is suggesting new testable hypotheses (Zednik & Jäkel, 2016) or measuring human performance against given standards of rationality (Crupi & Girotto, 2014; Tauber et al., 2017). Moreover, a certain degree of freedom in accommodating the observed behavioural data, as well as the tendency to be influenced by some kind of confirmation bias (Lakatosian "dogmatism"), are not epistemic peculiarities of Bayesian models but characterize any scientific account of the human mind (Griffiths et al., 2012b; Zednik & Jäkel, 2016). As observed by Griffiths and colleagues (2012b, p. 416), for instance, «Bayesian models are […] falsifiable as any empirical hypothesis – any hypothesis can be "saved" by suitable ad hoc adjustments to other aspects of the theory». Similarly, according to the same authors (*ivi*):

> [b]eing careful about the degrees of freedom and using appropriate procedures for comparing and testing models are important things to keep in mind for all forms of computational modelling; they do not constitute a problem that is specific to Bayesian models.

What is specifically problematic to certain applications of Bayesian modelling, as we have seen, is rather the peculiar combination of theoretical moves that characterizes the inferential interpretation and use of RA, a combination which has not been fully captured by standard criticisms of Bayesian models in cognitive science.[9]

On the other hand, some remarks put forward by advocates of Bayesian cognitive models are coherently vindicated in the light of what we called the constructive interpretation of RA. For instance, against the claim that Bayesian models make unwarranted claims about human rationality, it has been argued that the «Bayesian framework is a means of generating empirical hypotheses, rather than an assertion that people are optimal» (Griffiths et al., 2012b, p. 416). Similarly, some scholars have argued that «most of the heuristics that contribute to Bayesian reverse-engineering serve to formulate testable hypotheses, but not to directly support the claim that one of these hypotheses is actually true» (Zednik & Jäkel, 2016, p. 3973–74). In this sense, Bayesian models are just «catalysts for inspiration» (*ivi.* p. 3974), allowing «merely to navigate the space of computational-level hypotheses» (*ivi,* p. 3980) and thus making «an invaluable contribution to scientific discovery» (*ivi,* p. 3974). These remarks fit well with the spirit of the positive heuristic as characterized by Lakatos, of which the constructive variant of RA is but an instance, as we suggest.

[9] For instance, as we have claimed, the questionable move in inferential RA is not accommodation per se but the tacit equation of data accommodation with empirical support. This move is only loosely captured by saying that «theorists take the successful predictions of a Bayesian model as support to their approach and ignore the fact that alternative non-Bayesian theories might account for the data just as well, and sometimes better» (Bowers & Davis, 2012a, p. 403; see also Bowers & Davis, 2012b), as critics of Bayesian models have sometimes argued. In addition, as we have seen, in inferential RA rationality considerations are invoked with the specific purpose of providing support for empirical models, leading to a specific confusion between normative and descriptive considerations. Again, this confusion is only loosely captured by saying that «it is not always obvious when a Bayesian model is intended to imply a normative claim and when it is not» (Tauber et al., 2017, p. 411), as some scholars have observed.

8 Resource-RA

As a possible objection to the considerations that we have provided so far, one might argue that the inferential interpretation of RA might have been common only in the early phase of Bayesian cognitive science. In recent years – so the objection goes – normative considerations are progressively less relevant and theorists in this field are increasingly moving towards more realistic and descriptively adequate models of cognitive phenomena.

As a representative example, Lieder and Griffiths (2020) have recently formulated a modification of the methodology proposed by Anderson called *resource-rational analysis* (resource-RA), which gives more emphasis to computational limitations and the trade-off between optimality and psychological constraints. Here is a sketch of how it operates. To begin with, (i) the theorist formulates a computational-level theory of a cognitive problem and (ii) explicitly considers the class of algorithms that the mind might use to solve it as well as the cognitive costs of these algorithms. Then, (iii) an algorithm in this class that optimally trades off resources and approximation accuracy is individuated, and (iv) predictions made by the model are evaluated in light of behavioural data. If the predictions are off, (v) the theorist is expected to reiterate the analysis by refining either the computational-level theory or the assumed constraints. Alternatively, if the proposed cognitive model is already sufficiently realistic, the theorist can stop the analysis.

Resource-RA is explicitly presented as a constructive procedure for Bayesian cognitive science: «it provides a tool for replacing the traditional method of developing cognitive process models […] with a means of automatically deriving hypotheses about cognitive processes from the problem people have to solve and the resources they have available to do so» (Lieder & Griffiths, 2020, p. 6). Nevertheless, worries remain about the ability of resource-RA to overcome some limitations of the traditional approach, such as the reliance on a weak notion of evidential support from the data (point 4 above). As noted by Lieder and Griffiths themselves (p. 14),

> [e]ncouraging modellers to revise their assumptions about cognitive constraints in the face of data […] could also produce bad models that overfit observations of idiosyncratic or genuinely irrational behaviours with wrong assumptions.

The ability of resource-RA to avoid subtle violations of the *is / ought* divide is also yet to be assessed, provided that the appeal to rationality (under constraints) is still one of the avowed virtues and the central guiding principles of the approach.

Perhaps the most pressing worry of resource-RA is that it remains somehow anchored to the claim that humans are essentially rational (point 5). After all, the spirit of the project is to demonstrate that the reasoning mechanisms that are commonly interpreted as evidence against human rationality can be reinterpreted as reflecting the optimal use of finite time and limited computational resources (Lieder & Griffiths, 2020, p. 7). This claim appears to be inconsistent, however, with the established body of evidence supporting the idea that «computational limits don't fully explain human cognitive limitations» (Davis & Marcus, 2020, p. 21), given the existence of «cases [that] are not only suboptimal, but rather "*anti*-Bayesian", for actively defying Bayesian norms of inference» (Mandelbaum et al., 2020, p. 31). In light of the amount of data showing that human cognition is not rational or optimal, and not even boundedly rational under constraints, some Bayesian cognitive scientists have started to suggest «setting optimality aside and letting data drive psychological theory» (Tauber et al., 2017, p. 410).

9 Second exhibit: The conjunction fallacy and Bayesian confirmation theory

In order to figure out one way for a Bayesian perspective on cognition to possibly survive without being tainted by the methodologically dubious implications of inferential rational analysis, we will now refer to another widely know example from the psychology of reasoning. In certain circumstances, people have a systematic tendency to assess a conjunctive statement as more likely than one of its conjuncts, contrary to the principles of probability theory. A number of studies have documented this phenomenon, so-called "conjunction fallacy" (see Wedell & Moro, 2008; Tentori & Crupi, 2012 for an assessment). The most widely known illustration is of course the Linda scenario, taken from the seminal works of Tversky and Kahneman (1982, 1983). When faced with the description of a character, Linda (who is 31 years old, single, outspoken, and very bright, with a major in philosophy, concerns about discrimination and social justice, and an involvement in anti-nuclear demonstrations as a university student), most people ranked the statement "Linda is a bank teller and is active in the feminist movement" as more probable than "Linda is a bank teller".

The conjunction fallacy does not occur indiscriminately. *When* does it occur then? Pretty simple question to ask, not so easy to answer, it turns out. Tversky and Kahneman drew from their general framework for the study of judgment under uncertainty (Tversky & Kahneman, 1974), but they did not achieve (nor they claimed) full success in explaining the rich variety of their findings. Virtually all later approaches posited that the extent of the conjunction fallacy effect should simply increase with the judged probability of the added conjunct ("feminist", in the Linda scenario; see Tentori et al., 2013 for an overview).

If you are a Bayesian, you may wonder if there's a better interpretation of this phenomenon. Maybe some latent subtlety is implied when people face conjunction fallacy tasks. You notice that the "feminist" conjunct is not just relatively probable as concerned Linda; perhaps more interestingly, it clearly seems to be *confirmed* by the evidence initially provided (Linda's description) in the specific sense of Bayesian confirmation theory (Crupi et al., 2008). Indeed, one can account for all main extant variants of the phenomenon on the basis of confirmation-theoretic connections among the key elements in the scenario, the idea being — roughly — that people tend to make a conjunction fallacy judgment to the extent the conjunction is more strongly confirmed than the target isolated conjunct by available evidence that is explicitly given or psychologically salient (see Tentori et al., 2013; Crupi & Tentori, 2016). In this view, subtle (and sound) assessments of evidential relevance can guide judgments of probability, thus generating systematic biases (also see Tentori et al., 2016).

Here, one is letting the data "drive psychological theory" (or better, theory assessment) not because Bayesian theorizing is avoided, but because empirical support is not claimed from inferential RA. Indeed, some *novel, independent, and successful predictions* arise from this approach. For instance, in the Linda scenario, "Linda is a bank teller and a feminist activist" will be the target of *more* (not less) conjunction fallacy judgments than "Linda is a bank teller and owns a pair of black shoes", even if people appreciate that the "black shoes" conjunct is more probable than "feminist", because they also perceive that the former gets less evidential support than the latter from the information given (see Tentori et al., 2013). Even more surprisingly, perhaps, the confirmation-theoretic analysis implies that a substantial amount of *double* conjunction fallacy judgments can be obtained in case a conjunction is confirmed by the evidence while none of the conjuncts is — a pattern that cannot be anticipated by any competing account (see Crupi et al., 2018a).

10 Discussion

Let us recap. Our outline of the inferential form of RA indicates that (1)-(5) are not independent features of PBCS. Roughly, (1)-(4) generate premises by which the claim of human rationality can be derived (if informally). To repeat, a Bayesian model is developed (1), which is assumed to fulfil compelling rationality requirements (2); a weak notion of evidential support is invoked to claim empirical indistinguishability of Bayesian and alternative accounts (3), and the normative status of the Bayesian analysis is employed as a theoretical virtue to solve the problem of comparative assessment (4). Of course, in such context, the final acceptance of a Bayesian analysis entails the rationality of behavior (5).

But the claim of rationality is untenable. Despite the sustained effort, the epic task of rationalizing human cognition through PBCS has met only limited success. "Given the reams of evidence that cognition is fallible, [pure] Bayesians are fighting an uphill battle" (Fernbach & Sloman, 2011, p. 199). This raises the question of whether a Bayesian approach to cognition is possible at all, without the debatable features represented by (3), (4), and (5). Here, we take our second exhibit to provide an existence proof. For the sake of convenience, we will label this (by now admittedly minor) variant *critical* Bayesian cognitive science (CBCS), and characterize it as follows.

(1) It draws on the Bayesian framework and formal machinery to account for cognitive phenomena documented by behavioral data.
(2) It regards Bayesianism as a compelling general model of rational inference and judgment under uncertainty.
(3) *It meets the standards of a robust, predictivist view of evidential support from the data.
(4) *It does not imply any violation of so-called *is* / *ought* divide.
(5) *It acknowledges that some systematic violations of rational constraints in humans are real.

Note that (1) still distinguishes CBCS from the heuristic and biases approach in its traditional outlook / variation (Gilovich et al., 2002), while both (1) and (2) distinguish CBCS from the "fast and frugal heuristics" program by Gigerenzer and colleagues (Gigerenzer et al., 2011). We happen to consider (3*)-(5*) much more attractive than their counterparts (3)-(5) for both philosophical and scientific reasons. Someone who shares such inclinations may find comfort in the conclusion that we do *not* have to "throw out the Bayes with the bathwater" (Fernbach & Sloman, 2011).

It is not our aim to reconstruct the philosophy of science foundations of CBCS in detail in the present context. Note, however, that within cognitive science a similar perspective has been recently suggested by Tauber and colleagues (2017) under the name of *descriptive* Bayesian approach.[10] Within the descriptive approach outlined by Tauber and colleagues, a Bayesian model needs not to imply any claim that the underlying cognition is optimal or rational (even under constraints), and is used solely as a tool for building a psychological theory. This is a «vision that explicitly, as part of its structure, rejects assumptions or interpretations of optimality» (p. 413). In particular, in one of the case studies presented in the article (*case 1*), the authors formulate a Bayesian model that squarely incorporates some assumptions that are problematic from a normative point of view (such as inferential "conservatism", see pp. 418–22). Of course, as also Tauber and colleagues suggest, more work is needed to make this methodological approach take root in actual experimental practice. Nevertheless, we take this as a sign that the time is ripe for promoting a substantial change in perspective within (Bayesian) cognitive science, and we hope that the philosophical considerations provided here might serve the cause.

Acknowledgements Previous versions of this work have been presented at the Ruhr-Universität Bochum (2019), at a Harvard / Northeastern University workshop on *Bayesianism and its rivals* (2022), and at the 2022 Conference of the Philosophy of Science Association in Pittsburgh. We thank the audience for valuable comments, Peter Brössel for a particularly fruitful early exchange on these ideas, and two anonymous reviewers of EJPS for useful and constructive observations.

Author contribution Vincenzo Crupi has originally conceived the content of the article and written Sections 2–5 and 9–10. Fabrizio Calzavarini has helped revising the content and written Sections 6–8.

Funding Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement.

Declarations

Ethical approval Not applicable.

Informed consent Not applicable.

[10] Tauber and colleagues explicitly distinguish the descriptive approach from what they call the *normative* approach, which «uses a Bayesian model as a normative standard upon which to license a claim about optimality» (Tauber et al., 2017, p. 410). Note however that the normative approach as outlined by Tauber and colleagues appears to be different from the inferential use of RA as presented in this article. For the former approach is characterized just as the practice of using a Bayesian model as a normative standard against which human performance is measured, with no claim involved about the actual implementation of such model by human cognition (see Tauber et al., 2017, p. 411).

Financial or non-financial interests None.

References

Achourioti, T., Fugard, A. J., & Stenning, K. (2014). The empirical study of norms is just what we are missing. *Frontiers in Psychology, 5*, 1159. https://doi.org/10.3389/fpsyg.2014.01159

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Erlbaum.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14*, 471–485.

Barnes, E.C. (2022). Prediction versus accommodation. In E.N. Zalta and U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition). URL: https://plato.stanford.edu/archives/win2022/entries/prediction-accommodation.

Baron, J. (1991). Some thinking is irrational. *Behavioral and Brain Sciences, 14*, 486–487.

Baron, J. (2012). The point of normative models in judgment and decision making. *Frontiers of Psychology, 3*, 577.

Bowers, J. S., & Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*, 389–414.

Bowers, J. S., & Davis, C. J. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin, 138*, 423–426.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews (Cognitive Science), 1*, 811–823.

Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *British Journal for the Philosophy of Science, 68*, 451–484.

Colombo, M., & Seriès, P. (2012). Bayes in the brain—On Bayesian modelling in neuroscience. *British Journal for the Philosophy of Science, 63*(3), 697–723.

Colombo, M., Elkin, L., & Hartmann, S. (2021). Being Realist about Bayes, and the Predictive Processing Theory of Mind. *British Journal for the Philosophy of Science, 72*, 185–220.

Crupi, V., & Girotto, V. (2014). From *is* to *ought*, and back: How normative concerns foster progress in reasoning research. *Frontiers of Psychology, 5*, 219.

Crupi, V., & Tentori, K. (2016). Noisy probability judgment, the conjunction fallacy, and rationality: Comment on Costello and Watts (2014). *Psychological Review, 123*, 97–102.

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking and Reasoning, 14*, 182–199.

Crupi, V., Elia, F., Aprà, F., & Tentori, K. (2018a). Double conjunction fallacies in physicians' probability judgment. *Medical Decision Making, 38*, 756–760.

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

Crupi, V. (2020). Confirmation. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/confirmation/.

Crupi, V. (2023). The case of early Copernicanism: Epistemic luck *vs*. predictivis vindication (manuscript submitted).

Davis, E. S., & Marcus, G. F. (2020). Computational limits don't fully explain human cognitive limitations. *Behavioral and Brain Sciences, 43*, e7.

Dubey, R., & Griffiths, T. L. (2020). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review, 127*, 455–476.

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines, 21*, 389–410.

Elqayam, S., & Evans St, J. B. T. (2011). Subtracting *ought* from *is*. *Behavioral and Brain Sciences, 34*, 233–248.

Elqayam, S., & Over, D. E. (2016). From *is* to *ought*: The place of normative models in the study of human thought. *Frontiers of Psychology, 7*, 628.

Fernbach, P. M., & Sloman, S. A. (2011). Don't throw out the Bayes with the bathwater. *Behavioral and Brain Sciences, 34*, 198–199.

Fitelson, B., & Hawthorne, J. (2010). The Wason task(s) and the paradox of confirmation. *Philosophical Perspectives, 24*(Epistemology), 207–241.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98*, 254–267.

Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The Foundations of Adaptive Behavior*. Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Glymour, C. (2007). Bayesian Ptolemaic psychology. In W. Harper and G. Wheeler (eds.), *Essays in Honor of Henry E. Kyburg Jr.* (pp. 123–141). King's College Publishers.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*, 108–154.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116*, 661–716.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 59–100). Cambridge University Press.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012a). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21*, 263–268.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012b). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin, 138*, 415–422.

Hahn, U. (2014). The Bayesian boom: Good thing or bad? *Frontiers in Psyhcology, 5*, 765.

Heit, E. (1998). A Bayesian analysis of some forms of induction. In M. Oaksford & N. Chater, *Rational Models of Cognition* (pp. 248–274). Oxford University Press.

Icard, T. F. (2018). Bayes, bounds, and rational analysis. *Philosophy of Science, 85*, 79–101.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*, 169–231.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*, 307–321.

Kemp, C., Tenenbaum, J.B. (2003). Theory-based induction. *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 658–663). Psychology Press.

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes*. Cambridge University Press.

Laudan, L. (1990). Demystifying underdetermination. Scientific TheoriesIn C. W. Savage (Ed.), *Minnesota Studies in the Philosophy of Science* (Vol. 14, pp. 267–297). University of Minnesota Press.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43*, e1.

Mandelbaum, E. (2019). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Language, 2019*(34), 141–157.

Mandelbaum, E., Won, I., Gross, S., & Firestone, C. (2020). Can resources save rationality? "Anti-Bayesian" updating in cognition and perception. *Behavioral and Brain Sciences, 43*, e16.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher level cognition? *Psychological Science, 24*, 2351–2360.

Marcus, G. F., & Davis, E. (2015). Still searching for principles: A response to Goodman et al. (2015). *Psychological Science, 26*, 542–544.

Nickerson, R. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking and Reasoning, 2*, 1–31.

Oaksford, M. (2014). Normativity, interpretation, and Bayesian models. *Frontiers of Psychology, 5*, 332.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608–631.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review, 10*, 289–318.

Oaksford, M., & Chater, N. (2007). *Bayesian Rationality*. Oxford University Press.

Oaksford, M., & Chater, N. (2009). Précis of *Bayesian Rationality*: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences, 32*, 69–120.

Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses. *Psychological Bulletin, 144*, 779–796.

Rescorla, M. (2015). Bayesian perceptual psychology. In M. Matthen (Ed.), *The Oxford Handbook of Philosophy of Perception* (pp. 694–716). Oxford University Press.

Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind & Language, 31*, 3–36.

Sanjana N.E. and Tenenbaum J.B. (2003). Bayesian models of inductive generalization. *Advances in Neural Information Processing Systems* (pp. 59–66), MIT Press.

Sloman, S. A., & Fernbach, P. M. (2008). The value of rational analysis: An assessment of causal reasoning and learning. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 486–500). Oxford University Press.

Stein, E. (1996). *Without good reasons: The rationality debate in philosophy and cognitive science*. Clarendon Press.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review, 124*, 410–441.

Tenenbaum J.B. and Griffiths T.L. (2001). The rational basis of representativeness. In J. Moore and K. Stenning (eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 1036–1041). Erlbaum.

Tentori, K., & Crupi, V. (2012). On the conjunction fallacy and the meaning of and yet again: A reply to Hertwig, Benz, and Krauss (2008). *Cognition, 122*, 123–134.

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General, 142*, 235–255.

Tentori, K., Chater, N., & Crupi, V. (2016). Judging the probability of hypotheses vs. the impact of evidence: Which form of inductive inference is more accurate and time-consistent? *Cognitive Science, 40*, 758–778.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional *vs*. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293–315.

Vindrola, F., & Crupi, V. (2021). Bayesian too should follow Wason: A comprehensive accuracy-based analysis of the selection task. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/716170

Wason, P. (1966). Reasoning. In B. Foss (Ed.), *New Horizons in Psychology* (pp. 135–151). Penguin.

Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus and problem type. *Cognition, 107*, 105–136.

Worrall, J. (2011). Underdetermination, realism, and empirical equivalence. *Synthese, 180*, 157–172.

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese, 193*, 3951–3985.