THE WASON TASK(S) AND THE PARADOX OF CONFIRMATION

BRANDEN FITELSON AND JAMES HAWTHORNE

ABSTRACT. The (recent, Bayesian) cognitive science literature on the Wason Task (WT) has been modeled largely after the (not-so-recent, Bayesian) philosophy of science literature on the Paradox of Confirmation (POC). In this paper, we apply some insights from more recent Bayesian approaches to the (POC) to analogous models of (WT). This involves, first, retracing the history of the (POC), and, then, re-examining the (WT) with these historico-philosophical insights in mind.

1. THE PARADOX OF CONFIRMATION

Before getting into the recent (Bayesian) cognitive science literature on the Wason Task, we will re-trace the historical (and philosophical) background of the Paradox of Confirmation. In this section, we will tell a revised and extended version of a story about (the history and philosophy of) The Paradox that that we have (partially) told elsewhere [6, 8, 7]. In section two, we move on to the Wason Task(s).

1.1. Hempel and Goodman on the Paradox of Confirmation.

Not surprisingly, the Paradox of Confirmation involves a relation called *the confirmation relation*. For Hempel and Goodman (and also for Carnap, who'll enter our story shortly), the confirmation relation was a *logical* relation, akin to deductive entailment. Thus, in the heady early days of confirmation theory, ⌜E confirms H⌝ was meant to express a logical relation between propositions (or, better still, sentences) E and H. The precise nature of this logical relation will, of course, vary, depending on one's favorite explication of said logical confirmation concept. But, the basic idea (common to all such explications) is that "confirms" is supposed to be a *weaker* relation than "entails", and there are supposed to be substantive and non-trivial true instances of ⌜E confirms H⌝, where ⌜E entails H⌝ is *false*. So, on this traditional view, confirmation is a *generalization of (classical) entailment*.

In the beginning, there was Hempel's confirmation relation [16], which is purely syntactical (defined over a first-order language $\mathcal{L}$), and which fully supervenes on classical deductive entailment relations (of $\mathcal{L}$). The full details of Hempel's confirmation theory won't be crucial for present purposes. We'll only need to know a few of the formal consequences of Hempel's theory — mainly, the following two:

(NC) For all names $\mathbf{x}$ and for all (classically) logically independent predicate expressions ϕ and ψ: ⌜$\phi\mathbf{x} \,\&\, \psi\mathbf{x}$⌝ confirms ⌜$(\forall \mathbf{y})(\phi\mathbf{y} \supset \psi\mathbf{y})$⌝.

(EQC) For all statements, E, H, and H': if E confirms H and H is (classically) logically equivalent to H', then E also confirms H'.

Date: 08/01/10. *Draft*: Final version to appear in *Philosophical Perspectives*, J. Hawthorne (*ed.*).

Informally, (NC) asserts that "universal laws are confirmed by their positive instances,"[1] and (EQC) presupposes that confirmation relations depend only on the logical content (in a naive, classical sense) of the statements involved. If confirmation is a logical relation that generalizes classical entailment, then (EQC) should be sacrosanct.[2] On the other hand, (NC) is far more controversial (no matter how we think of the confirmation relation). Indeed, as we'll see, (NC) is perhaps *the* crucial underlying principle both in the context of the Paradox, and in the context of the Wason Task. So, we'll have a lot more to say about (NC) as the paper progresses. Meanwhile, let's return to the Paradox of Confirmation itself.

The Paradox of Confirmation can be generated from the two basic consequences of Hempel's theory of confirmation: (NC) and (EQC), *via* the following very simple argument. Let $Rx \stackrel{\text{def}}{=} x$ is a raven, and $Bx \stackrel{\text{def}}{=} x$ is black. Then, we have:

(1) '$\sim Ba \,\&\, \sim Ra$' confirms '$(\forall x)(\sim Bx \supset \sim Rx)$'. [instance of (NC)]
(2) '$(\forall x)(\sim Bx \supset \sim Rx)$' is equivalent to '$(\forall x)(Rx \supset Bx)$'. [classical logic]
(PC) $\therefore$ '$\sim Ba \,\&\, \sim Ra$' confirms '$(\forall x)(Rx \supset Bx)$'. [(1), (2), (EQC), logic]

(PC) — the so-called "Paradoxical Conclusion" — says that the statement that a is a non-black non-raven confirms the statement that all ravens are black. This was thought by many to have a "paradoxical ring" to it. Goodman [12, p. 71] famously quipped that (PC) would sanction "indoor ornithology", since (PC) seems to suggest that observing (*e.g.*) white shoes can allow us to gather evidence that is relevant to hypotheses about ravens. That rhetorical flourish is misleading, for two reasons.

First, all that follows from (NC) and (EQC) is that '$\sim Ba \,\&\, \sim Ra$' confirms that all ravens are black. It does *not* follow from (NC) and (EQC) that (*e.g.*) '$Wa \,\&\, Sa$' confirms that all ravens are black (where $Wx \stackrel{\text{def}}{=} x$ is white, and $Sx \stackrel{\text{def}}{=} x$ is a shoe), even if we stipulate that Wa entails $\sim Ba$ and Sa entails $\sim Ra$. Interestingly, Hempel's *full* theory *does* have this "Goodmanian" consequence, since it entails:

(M) For all names $\mathbf{x}$, for all (classically) consistent predicates ϕ and ψ, and for all statements H: If ⌜$\phi\mathbf{x}$⌝ confirms H, then ⌜$\phi\mathbf{x} \,\&\, \psi\mathbf{x}$⌝ confirms H.

(M) is a a kind of *evidential monotonicity* property. What (M) says is that if some statement about an object confirms some hypothesis, then any logically stronger statement about said object must also confirm said hypothesis. Monotonicity is a surprising property for an *inductive*-logical relation to have. Indeed, it's one which conflicts with *Hempel's own intuitions* about inductive logic! We'll return to this interesting (and overlooked) historical anomaly in the work of Hempel (and Goodman) shortly. But, for now, we just want to point out that (M) is presupposed by Hempel's theory, and also by Goodman's infamous "indoor ornithology" remark.

Second, even if (PC) *did* imply that statements about white shoes confirm the hypothesis that all ravens are black, it's still not clear why that should be *problematic* for a *logical* confirmation relation. Here's an analogy. According to classical

[1] This presupposes that, *e.g.*, '$Ra \,\&\, Ba$' is a "positive instance" of '$(\forall x)(Rx \supset Bx)$'. That will strike some readers as odd. In modern logic, we would think of '$Ra \supset Ba$', rather than '$Ra \,\&\, Ba$', as an instance of '$(\forall x)(Rx \supset Bx)$'. This (odd) way of thinking about "instances" traces back to Keynes [21, Ch. XIX]. While this difference makes no difference in the context of Hempelian confirmation theory, it makes a big difference in the context of Bayesian confirmation theory [7, §3.4.1].

[2] We won't worry about (EQC) in this paper, although one could even question it on logical grounds, if one had worries about classical logic [36]. Moreover, one *should* question (EQC) if confirmation is an *epistemic* relation, since then (EQC) becomes a substantive claim concerning the normativity of logic [13]. We will only briefly touch on those sorts of issues here. See [7] for further discussion.

logic, everything follows from any inconsistent set of statements. This is sometimes called "explosion". One might think that explosion is problematic, on the grounds that it would sanction arbitrary *inferences* from inconsistent sets of *beliefs*. But, as Harman [13] rightly points out, explosion is a logical rule, and logic (*alone*) doesn't tell us which inferences are kosher and which are not. That's an *epistemological* question, not a logical one. So, the fact that everything classically follows from an inconsistent set of statements is not (by itself) a reason to reject classical logic *qua logic*. In order to parlay this into an argument against classical logic, one would need some sort of *bridge principle* to connect logic and epistemology (*i.e.*, entailment and inference). And, if one reflects on the kinds of bridge principles one would need here, one quickly comes to realize that this is a much more difficult task than one might have initially thought.[3] Analogously, the fact that Hempel's logical relation of confirmation has the property that '$\sim Ba \,\&\, \sim Ra$' (or even '$Wa \,\&\, Sa$') confirms '$(\forall x)(Rx \supset Bx)$' is only problematic to the extent that we conflate this *logical* claim with some *epistemic* claim like "observing white shoes is a way of obtaining evidence that is relevant to the claim that all ravens are black". To be sure, it is something like this epistemic claim about *evidential support* that underlies Goodman's "indoor ornithology" worry, and the (alleged) paradoxicality of The Paradox. Unfortunately, Hempel and Goodman never discussed any precise bridge principles of the requisite sort. But, they were (implicitly) sensitive to some of the thorny underlying philosophical issues here. We can see this most clearly in their initial response to The Paradox itself, to which we now turn.

Because Hempel's theory entails (PC), and Hempel was well aware of this entailment, he recognized the need to address its (apparent) paradoxicality. Officially, Hempel (and Goodman) took the position that (PC) was not really paradoxical at all. And, Hempel (and Goodman) endorsed the same sort of strategy for "explaining away" the appearance of a paradox. Hempel [17, $p.$ 20] explains that

> … in the seemingly paradoxical cases of confirmation, we are often not judging the relation of the given evidence E *alone* to the hypothesis H … instead, we tacitly introduce a comparison of H with … E in conjunction with … additional … information we … have at our disposal.[4]

What Hempel is suggesting here is that confirmation should be thought of as a three-place relation: 'E confirms H, relative to a background corpus K'. And, that the paradoxical conclusion only (properly) holds when K is *tautological* ($\top$) or *empty*. Specifically, Hempel is warning us not to conflate the following two claims:

(PC) $\sim Ba \,\&\, \sim Ra$ confirms $(\forall x)(Rx \supset Bx)$, *relative to* $\top$.
(PC*) $\sim Ba \,\&\, \sim Ra$ confirms $(\forall x)(Rx \supset Bx)$, *relative to* $\sim Ra$.

Hempel thought that, *intuitively*, (PC) is true, but (PC*) is false. That is, he had the intuition that *if* K already contains the information that a is a non-raven (*i.e.*, if $K \vDash \sim Ra$), *then* $\sim Ba \,\&\, \sim Ra$ does *not* confirm $(\forall x)(Rx \supset Bx)$, relative to K. But, if $K = \top$, then $\sim Ba \,\&\, \sim Ra$ *does* confirm $(\forall x)(Rx \supset Bx)$, relative to K. Hempel suggests that people who find (PC) unintuitive are conflating (PC) with (PC*), and this is why they are (mis)lead to suspect that (PC) is false. Throughout Hempel's

[3] We borrow the term "bridge principle" from John MacFarlane [22]. See [7] for a discussion of this Harmanian point (as applied to *inductive* logic), and its ramifications for Goodman's "grue" paradox.

[4] Goodman [12, $p.$ 70] offers a very similar "explaining away" of the apparent paradoxicality of (PC). Goodman [12, $p.$ 71] also endorsed an independent defense of (PC), which was pioneered by Israel Scheffler [34, $pp.$ 186–191], and revisited later in a paper Goodman wrote with Scheffler [35].

(and Goodman's) "explaining away", we see vacillation between *logical* readings of "confirms" and *epistemic* readings of "confirms". This makes it somewhat difficult to precisely pin down the crux of their "explaining away" strategy. The best reconstruction of the core of the Hempel/Goodman "explaining away" of the Paradox that we have been able to come up with is the following central *epistemic intuition*:

($\mathscr{E}$) If S already knows that a is a non-raven, then S's observing a's color will not generate any evidence (for S) about the color of ravens. But, if S knows nothing about a, then S's learning (say, by observation of a) that $\sim Ba \& \sim Ra$ does provide some evidence (for S) that all ravens are black.

Hempel and Goodman never actually explicitly state anything as clear as ($\mathscr{E}$). Nor do they ever give a compelling argument for anything like ($\mathscr{E}$). We are employing a bit of charity (and hindsight) here. While we're at it, allow us to provide a rationale for ($\mathscr{E}$), which strikes us as pretty reasonable [23]. If one learns $\sim Ra$ (or Ba[5]), then one is learning something which entails that a is not a counterexample to the law "all ravens are black" (H). In this sense, learning $\sim Ra$ (or Ba) has the consequence of *ruling-out a possible counterexample to* H. This elimination of a possible counterexample to H has the effect of reducing the size of the set of possible counterexamples to H (by one); and, thereby, *indirectly lending (some) support to the truth of* H. If one buys this "indirect support" story, then one has some reason to accept ($\mathscr{E}$). The idea behind the first part of ($\mathscr{E}$) is that if one *already knows* $\sim Ra$, then the "confirmational boost" for H has *already* occurred, and so learning (in addition) that $\sim Ba$ adds no *further* support to H. The second part of ($\mathscr{E}$) is straightforward, since that just says that — if one starts out with *no* knowledge about a — learning $\sim Ba \& \sim Ra$ lends some support to H, because this too involves learning something which entails that a is not a counterexample to H. Once we understand the Hempelian/Goodmanian strategy in these (epistemic) terms, we can see that it's a pretty reasonable (and even clever) strategy for "explaining away" the apparent (epistemic) paradoxicality of (PC).

However, without some (inductive) bridge principle to connect the *logical* concept of confirmation and the *epistemic* concept of evidential support, it is rather mysterious why (PC) should seem *logically* (*viz.*, *confirmation*-theoretically) paradoxical in the first place, since it's unclear what ($\mathscr{E}$) has to do with *logic* (*viz.*, *confirmation*). For this reason, we will (charitably) assume that Hempel and Goodman are (tacitly) presupposing something like the following (naive) bridge principle:

(BP) E evidentially supports H for S (in a context C) iff E confirms H, relative to K, where K is S's total evidence in context C.[6]

[5] Intuitively, observing the color of a known non-raven *won't tell you anything* about the color of ravens. On the other hand, the most one can say about the observation of the species of a known black object is that it *cannot refute* the claim that all ravens are black. So, while this "indirect inductive support" argument seems to apply equally to (antecedent) knowledge of $\sim Ra$ or Ba, there does seem to be an asymmetry here which makes the $\sim Ra$ version of the argument more compelling overall.

[6] Later in the paper, we will be discussing *quantitative* (and comparative) confirmation theory, and its applications to epistemology. For those applications, we'll (tacitly) presuppose a quantitative analogue of this naive bridge principle (BP), which will be (something like) the following:

(BP′) E evidentially supports H *to degree* r for S (in a context C) iff E confirms H, relative to K, *to degree* r where K is S's total evidence in context C.

See [7] for discussion of qualitative and quantitative bridge principles like (BP)/(BP′).

This is (basically) a Carnapian "requirement of total evidence" [39, p. 189]. In our paper on "grue" [7], we discuss these sorts of principles in more detail, and we explain their centrality in the context of Goodman's "grue" argument against Carnapian (and Hempelian) confirmation theory. Such issues involving the normativity of inductive logic are fascinating, but they are beyond the scope of the present paper. We will only briefly touch on these sorts of issues as the paper progresses.

To simplify the rest of the paper, we will (from now on) use the term "confirms" in a way that is (intentionally) ambiguous between its logical and epistemic readings (as Hempel and Goodman did). That is, we will use the word "confirms" in a way that can be read either logically (as a generalization of entailment) or epistemically (as something like "evidentially supports"). Hopefully, context will disambiguate where necessary. And, in this spirit of Hempelian ambiguity, we will re-word principle ($\mathscr{E}$) so that it is couched purely in terms of "confirmation", as:

($\mathscr{E}$) If $K \vDash \sim Ra$, then $\sim Ba \,\&\, \sim Ra$ does *not* confirm $(\forall x)(Rx \supset Bx)$, *relative to* K. But, if $K = \top$, then $\sim Ba \,\&\, \sim Ra$ *confirms* $(\forall x)(Rx \supset Bx)$, *relative to* K.

Unfortunately, while Hempel's principle ($\mathscr{E}$) is not that implausible — from an *intuitive* point of view (especially, when read *epistemically*, as above) — it turns out that ($\mathscr{E}$) is *incompatible* with Hempel's *official theory* of confirmation. This incompatibility (which seems to have gone unnoticed by Hempel, Goodman, and all subsequent commentators on confirmation theory) is caused by (i) the fact that Hempel's theory entails *monotonicity* (M), and (ii) the fact that Hempel's theory of confirmation (which is defined in terms of entailment relations) has no way of distinguishing $\ulcorner E$ confirms H, relative to $K\urcorner$ and $\ulcorner E \,\&\, K$ confirms H, relative to $\top\urcorner$. As a result, Hempel's theory entails the following ternary variant of monotonicity:

(M*) For all names $\mathfrak{x}$, (classically) consistent predicates ϕ and ψ, and statements H: If $\ulcorner \phi\mathfrak{x}\urcorner$ confirms H relative to $\top$, then $\ulcorner \phi\mathfrak{x}\urcorner$ confirms H relative to $\ulcorner \psi\mathfrak{x}\urcorner$.

Of course, it follows straightway from (M*) that (PC) *entails* (PC*). Thus, Hempel's (*intuitive*) claim that (PC) is true but (PC*) is false [*viz.*, ($\mathscr{E}$)] is incompatible with his *theory* of confirmation. So, this "explaining away" of the Paradox is not theoretically available to Hempel. As a result, if we want to vindicate this (intuitive) Hempelian strategy for "explaining away" the apparent paradoxicality of (PC), we'll need a theory of confirmation that is (at the very least) *non-monotonic*.

One final point about Hempel and Goodman, before moving ahead in the history of the Paradox. There is an important omission in these early discussions of the Paradox, and that is a distinction between qualitative, comparative, and quantitative varieties of confirmation. One wants to be able to say things like $\ulcorner E_1$ confirms H *more strongly than* E_2 confirms $H\urcorner$, or $\ulcorner E$ confirms H *to a minute degree*$\urcorner$, *etc.* Specifically, perhaps $\sim Ba \& \sim Ra$ *does* confirm $(\forall x)(Rx \supset Bx)$, *but only to a minute degree*. Or, perhaps $\sim Ba \& \sim Ra$ confirms $(\forall x)(Rx \supset Bx)$ *less strongly than* $Ba \& Ra$ does. The early, qualitative theories of confirmation (like Hempel's) are *too coarse grained* to capture such assertions. Indeed, these sorts of claims (which will be central to contemporary Bayesian approaches to the Paradox) cannot even be expressed in Hempel's theory. This is another of its important shortcomings.[7]

[7]To be fair, Hempel & Oppenheim [19] did propose a *quantitative* (and comparative) theory of degree of confirmation. But, their theory is still too coarse-grained to capture the nuanced kinds of comparative claims that will be crucial for us. Carnap [1, Ch. VII] later tried to come up with a probabilistic reconstruction of (quantitative and comparative) Hempelian confirmation theory. We will return to Carnap's probabilistic reconstruction of Hempel's theory at the very end of this paper.

Contemporary approaches to confirmation (*e.g.*, Bayesian approaches) have various advantages over Hempel's theory of confirmation. The two most important of these advantages (for present purposes) are (a) non-monotonicity, and (b) the ability to explicate fine-grained quantitative and comparative concepts of confirmation. We will discuss various Bayesian approaches to the Paradox shortly. But, first, we will digress (briefly) to discuss Quine's approach to the Paradox of Confirmation.

1.2. **Quine on the Paradox of Confirmation.**

Quine [31] motivates his entire discussion of (and approach to) natural kinds *via* an analysis the Paradox of Confirmation. This is somewhat surprising, since one might have thought that the "grue" paradox [7] would have been a more appropriate starting point for such a discussion. The reason Quine takes The Paradox of Confirmation as his point of departure is that he thinks it is a simpler illustration of the (instantial) confirmation-theoretic ramifications of the occurrence of non-natural kind terms in universal generalizations. Quine thought that *non*-blackness and *non*-ravenhood were paradigm examples of *non-natural kinds*. In general, Quine thought that the complements (or negations) of natural kinds (or natural kind terms) were themselves (highly) *non*-natural.[8] This was the key to Quine's approach to The Paradox. Quine thought that (NC) is false, as stated. But, he thought that (NC) is true — *if its scope is restricted to natural kinds*. In other words, Quine's approach to The Paradox is to reject step (1) of the simple argument for (PC) above. Quine's alternative to (NC) is the following principle:

(NC′) For all names $\mathfrak{x}$ and for all (classically) logically independent predicate expressions ϕ and ψ: *If* ϕ *and* ψ *denote natural kinds*, then $\ulcorner \phi\mathfrak{x} \,\&\, \psi\mathfrak{x}\urcorner$ confirms $\ulcorner (\forall y)(\phi y \supset \psi y)\urcorner$.

Because Quine thought that $\sim R$ and $\sim B$ do not denote natural kinds, his (NC′) cannot be instantiated to yield step (1) of our simple argument for (PC). So, Quine's diagnosis of the source of The Paradox is the falsity of step (1) of our simple argument; and, more generally, the falsity of (NC) *if not restricted to natural kinds*.

While Quine's approach to the paradoxes of confirmation (both the "raven" paradox and the "grue" paradox) is simple and unified, we do not think it is (ultimately) probative. As we will see in the next section, Bayesian insights about confirmation will reveal that (NC) is false, but for reasons that are orthogonal to Quine's worries about "naturalness". Indeed, we will soon see compelling Bayesian-style counterexamples to both (NC) and (NC′). This brief digression on Quine is mainly of historical interest, as it is an interesting chapter in the checkered career of (NC).

1.3. **Probabilistic Approaches to The Paradox of Confirmation.**

Probabilists have more sophisticated formal explications of confirmation relations than Hempel did. Hempel's explications were based on entailment relations, whereas probabilistic explications are based on *conditional probabilistic relevance* relations. This leads to the following formal explication of *qualitative* confirmation:

Qualitative. E confirms H, relative to K iff $\Pr(H \mid E \,\&\, K) > \Pr(H \mid K)$.

[8]Interestingly, Goodman (and Hempel and Carnap) did not agree with Quine on this point, which is one reason they all thought that "grue" really was a *new* riddle of induction. We tend to side with Hempel, Goodman, and Carnap on this point, but we won't have space to discuss that here.

Here, $\Pr(\cdot \mid \cdot)$ is some "suitable" conditional probability function. Different historical figures will have different ideas about which conditional probability functions are "suitable" for various sorts of confirmation-theoretic purposes. For instance, Carnap [1] required that $\Pr(\cdot \mid \cdot)$ itself have a *logical* construction/interpretation. Over the course of around 30 years of work on inductive logic, Carnap proposed various formal constructions of $\Pr(\cdot \mid \cdot)$ for confirmation-theoretic purposes. We won't delve into the details of these many Carnapian explications of conditional probability (and their applications to confirmation theory). But, we will discuss some of the properties of his later probability models, in connection with the Paradox. Then, we will discuss the historical movement away from the Carnapian "logical" approach to confirmation, in favor of the more subjective/psychologistic Bayesian approaches of the modern era. This will lead to a discussion of contemporary Bayesian methods in cognitive science (as applied to the Wason Task).

However, before we get into the historical dialectic of probabilistic confirmation theory, The Paradox, and the Wason Task, we first need to talk about *quantitative* and *comparative* probabilistic notions of confirmation. Bayesians have quantitatively generalized their qualitative (probabilistic relevance) confirmation relation in a variety of ways. The basic idea involves the adoption of a *confirmation measure* $\mathfrak{c}(H, E \mid K)$, which gauges "the degree to which E is probabilistically (confirmationally) relevant, conditional on K". As it turns out, there are *many* distinct confirmation measures floating around the literature [4]. Happily, for present purposes, we won't need to worry about this plethora of competing measures.[9] Moreover, we won't even need to think about comparative confirmation in its full generality. Because of the logical structure of the Paradox of Confirmation (and the Wason Task), we'll only need to worry about comparisons of the form "E_1 confirms H (relative to K) more strongly than E_2 confirms H (relative to K)", which will be explicated as:

> **Comparative**. E_1 confirms H (relative to K) more strongly than E_2 confirms H (relative to K) iff $\mathfrak{c}(H, E_1 \mid K) > \mathfrak{c}(H, E_2 \mid K)$.

Finally, we can simplify our discussion even further, by restricting our attention to measures of confirmation $\mathfrak{c}$ with the following crucial property:

(†) $\mathfrak{c}(H, E_1 \mid K) > \mathfrak{c}(H, E_2 \mid K)$ iff $\Pr(H \mid E_1 \,\&\, K) > \Pr(H \mid E_2 \,\&\, K)$.[10]

That allows us to concern ourselves only with the following (restricted in scope, and simplified) way of thinking about comparative confirmation

> **Comparative**. E_1 confirms H (relative to K) more strongly than E_2 confirms H (relative to K) iff $\Pr(H \mid E_1 \,\&\, K) > \Pr(H \mid E_2 \,\&\, K)$.

In other words, if we want to determine whether E_1 confirms H (relative to K) more strongly than E_2 confirms H (relative to K), all we have to do is compare the conditional probabilities $\Pr(H \mid E_1 \,\&\, K)$ and $\Pr(H \mid E_2 \,\&\, K)$. Returning to the Paradox, we are now ready to discuss two kinds of probabilistic approaches to it.

[9] Toward the end of the paper, we will discuss some Bayesian models of the Wason Task that *can* be sensitive to choice of confirmation measure. But, we won't delve into that. We will focus on other, more important problems with the typical Bayesian approaches to the Paradox and the Wason Task.

[10] See [6] and [8] for discussions of this property (†) of confirmation measures, and why measures that *violate* this property are unsuitable for applications to the Paradox of Confirmation. Basically, any measure $\mathfrak{c}(H, E \mid K)$ that is a function of the posterior $[\Pr(H \mid E \,\&\, K)]$ and prior $[\Pr(H \mid K)]$ probabilities of H will inevitably satisfy property (†). But, there are various confirmation measures in the literature which are not functions of the posterior and prior of H, and which violate (†) as a result.

1.3.1. *Probabilistic Approaches to the Paradox I: Rejecting (NC).*

The first type of probabilistic response to the Paradox was proposed by I.J. Good [9]. Good's approach was similar to Quine's, in the sense that both Quine and Good thought that the problematic assumption underlying the Paradox was (NC). But, Good's reasons for rejecting (NC) were much different than Quine's. Good pointed out that (NC) is not generally true, from a *probabilistic relevance* perspective. He gave the following sort of "random sampling model" counterexample to (NC).

> Let K be: Exactly one of the following two hypotheses is true: (H) there are 100 black ravens, no nonblack ravens, and 1 million other things in the universe [*viz.*, $(\forall x)(Rx \supset Bx)$], or ($\sim H$) there are 1,000 black ravens, 1 white raven, and 1 million other things in the universe.
>
> Let E be $Ra \,\&\, Ba$ (with a randomly sampled from the universe). Then:
>
> $$\Pr(E \mid H \,\&\, K) = \frac{100}{1000100} \ll \frac{1000}{1001001} = \Pr(E \mid \sim H \,\&\, K)$$
>
> $\therefore$ E lowers the probability of (*viz.*, *dis*confirms) H, relative to K.

At this point, many probabilists thought the Paradox had been dissolved. After all, it seemed that the Paradox had been shown to rest on a false confirmation-theoretic principle (NC), which was a vestige of old, deductive (Hempelian) ways of thinking about confirmation. And, or so the story continued, once we avail ourselves of the more sophisticated probabilistic relevance concept, we can see that there is no paradox here at all. Sometimes "positive instances" confirm generalizations, and sometimes they don't. This all depends on the details of the (statistical) structure of the situation, as encoded in our (statistical) background knowledge K.

While that story may sound very compelling, all is not beer and skittles. Hempel, in his reply to Good [18], was not moved by Good's counterexample to (NC). Hempel reminded Good that (NC) and (PC) were only meant to be asserted *relative to tautological or empty background corpus*. Hempel already believed (long before Good's paper appeared) that, once we introduce substantive empirical (say, statistical) information into our background corpus, this can undermine confirmation relations that had obtained relative to empty background corpus. Indeed, that was the key to Hempel's distinction between (PC) and (PC*). Not surprisingly, Hempel makes a similar distinction between (NC) which is relative to empty K ($\top$) and (NC*) which is relative to non-empty K. Since Good's K is non-empty, Hempel's response is that Good's "counterexample" only shows that (NC*) is false, but it does not show that (NC) is false. And, since Hempel had already pointed out the intuitive falsity of claims like (NC*), he was not surprised to find that probabilistic counterexamples to (NC*) could be constructed. Unfortunately, neither Hempel nor Good realized that this dialectical move *was not open to Hempel*, because (owing to monotonicity, as we explained above), Hempel's *theory* of confirmation is *incompatible* with the claim that (PC)/(NC) are true, while (PC*)/(NC*) are false. Had Good known this, of course, he would have had a very simple and powerful rejoinder to Hempel's response. Instead, Good (sort of) answered Hempel's (NC)/(NC*) challenge with a different rejoinder [10], in which he asks us (only half facetiously) to imagine

> . . . an infinitely intelligent newborn baby having built-in neural circuits enabling him to deal with formal logic, English syntax, and subjective probability. He might argue, after defining a crow in detail, that it is initially extremely unlikely that there are any crows [$\therefore$] extremely likely that all crows are black [but] if there are crows, then there is a reasonable chance they are a variety of colours [$\therefore$] if he were to discover that a black crow exists he would consider [H] to be less probable than it was initially.

The reason Good's rejoinder to Hempel is not entirely serious is that Good — like most other contemporary probabilists and statisticians — doesn't really know what it *means* to talk about "the probability of H relative to *tautological* or *empty* background corpus". Hempel's challenge would require a probabilist to make sense of just this sort of locution, since their confirmation theory would require them to compare $\Pr(H \mid E)$ and $\Pr(H \mid \top)$, for some "suitable" conditional probability function $\Pr(\cdot \mid \cdot)$. This is the point in the historical dialectic at which confirmation theory starts to tear away from its "logical/Hempelian roots". In order to take Hempel's challenge seriously, probabilistic confirmation-theorists would have to first explicate "$\Pr(H \mid \top)$", before they could engage with Hempel on key confirmation-theoretic questions. And, most modern probabilists don't think that any such explication is forthcoming. One reason they are skeptical is that Carnap spent 30+ years trying to provide an adequate explication of "$\Pr(H \mid \top)$", and — even by his own lights — he was not fully successful. Before moving on to the modern, Bayesian approaches to confirmation, The Paradox, and The Wason Task, we will digress (briefly) to consider what a *Carnapian* might say about The Paradox.

There are very few philosophers around today who still believe that Carnap's project of searching for an explication of "$\Pr(H \mid \top)$" — for the purposes of constructing a "Hempel-friendly" probabilistic confirmation theory — isn't futile. Patrick Maher is one of those few. In a series of recent papers, Maher has attempted to revive Carnap's project of explicating "logical probability" [*viz.*, "$\Pr(H \mid \top)$"], and largely with an eye toward confirmation-theoretic applications. For our purposes, the most salient paper of Maher's is his 2004 paper [24], in which he presents what he takes to be a fully adequate explication of "$\Pr(H \mid \top)$" for monadic predicate-logical languages $\mathcal{L}$ containing two (families of) predicates. As it turns out, Maher's models (which we will call *Carnapian* models) can be neatly applied to The Paradox of Confirmation. Indeed, Maher does apply his Carnapian models to The Paradox, and with some very interesting results. We won't delve into the technical details here. Rather, we will just mention a few of the most important results. First, Maher shows that (NC) fails to hold (generally) — even in the class of Carnapian models. The formal Carnapian counterexamples to (NC) are not easily digestible. But, Maher provides the following informal characterization of one type of Carnapian (perhaps even *Hempelian?*) counterexample to (NC) that he constructs in his paper:

> According to standard logic, 'All unicorns are white' is true if there are no unicorns. Given what we know[11], it is almost certain that there are no unicorns and hence 'All unicorns are white' is almost certainly true. But now imagine that we discover a white unicorn; this astounding discovery would make it no longer so incredible that a non-white unicorn exists and hence would disconfirm [*lower* the probability of] 'All unicorns are white.'

This example is somewhat similar to Good's "super-baby" example, but it makes more of an effort to take seriously Hempel's challenge regarding (NC) *vs* (NC*). To

[11] Even this example is one in which the background knowledge of the agent has substantial empirical content. This is probably inevitable (to some extent), and it underscores just how difficult it is to give an *intuitively compelling* counterexample to *Hempel's* (NC). Maher (personal communication) explains that the Unicorn example is only compelling *in conjunction with* his *formal* models. He says:

> The unicorn example is there because it is intuitive and the (qualitative) structure of the probabilities in it is the same as in the situation with no prior evidence but a low *a priori* probability for something to be an F [where the law is $(\forall x)(Fx \supset Gx)$]. Once we see how the probabilities work in the unicorn example, we can see that in general, when the prior probability of something being an F is low, Nicod's condition (NC) can be violated.

our ears, this sounds like a pretty plausible counterexample, even to Hempel's (NC). We think the lesson here is that (NC) isn't generally true — from a probabilistic-relevance point of view — even if it is understood as relativized to "empty background corpus".[12] So, rejecting (NC) may be a viable way for a probabilist to respond to Hempel's Paradox after all. Of course, the viability of this response depends on the viability of Carnapian explications of "$\Pr(H \mid \top)$". Most modern probabilists have abandoned this type of response to The Paradox, because they don't believe there is any such thing as "$\Pr(H \mid \top)$".[13] Before moving on to the more modern (and more subjectivist) approaches of contemporary Bayesians, we will discuss one other feature of Maher's Carnapian confirmation-theoretic models.

Recall that Hempel's own response to the Paradox was to blame the apparent paradoxically of (PC) on the conflation of (PC) and (PC*). As we have explained, because Hempel's theory is monotonic (M), it is inconsistent with this (PC)/(PC*) distinction. But, what about Carnapian confirmation-theoretic models? Are they compatible with (PC) being true while (PC*) is false? Note that, in probabilistic terms, (PC) and (PC*) would have the following forms, where $H \triangleq (\forall x)(Rx \supset Bx)$:

(PC) $\Pr(H \mid {\sim}Ba \,\&\, {\sim}Ra) > \Pr(H \mid \top)$.
(PC*) $\Pr(H \mid {\sim}Ba \,\&\, {\sim}Ra) > \Pr(H \mid {\sim}Ra)$.

Interestingly, Maher's Carnapian models are such that *it is possible for (PC) to be true while (PC*) is false*. In other words, Maher's Carnapian models *are* compatible with Hempel's (PC)/(PC*) distinction. This is (ultimately) because of the requisite sort of *non-monotonicity* of probabilistic relevance relations (even Carnapian ones). We think this is an important virtue of probabilistic relevance approaches to confirmation. After all, Hempel's ($\mathscr{E}$)-based intuitions about (PC) and (PC*) are not altogether implausible. And, so, we think it is important that probabilistic approaches to The Paradox be *compatible* with this fundamental Hempelian distinction. We will return to this point in our discussion of Bayesian approaches to the Paradox.

To sum up this section: the first type of probabilistic response to the Paradox is to show that (NC) can fail — from a probabilistic relevance point of view. The early counterexamples to (NC) were not "Hempel-friendly", since they involved cases in which the background corpus contains lots of empirical (statistical) information. Good's attempt to give a "Hempel-friendly" ("super-baby") counterexample was only half-hearted. But, later, Maher was able to show that his Carnapian probability models can undergird much more interesting "Hempel-friendly" counterexamples to (NC). We also saw that (*pace* Quine) these Carnapian counterexamples to (NC) do not seem to (essentially) involve "non-natural" kinds (in Quine's sense). And, finally, we saw that (unlike Hempel's own theory) Maher's Carnapian models are compatible with Hempel's (PC)/(PC*) distinction. Unfortunately, however, Carnapian approaches to confirmation require the existence of "logical probabilities". And, most modern probabilists are quite skeptical about the existence (and/or probative value) of "$\Pr(H \mid \top)$". This is why almost all contemporary approaches to The

[12] And, even if (NC) is restricted to laws involving "natural kinds" [as Quine's (NC′)]. We don't mean to suggest here that "unicorn" is a natural kind. But, there doesn't seem to be anything *essentially "non-natural"* (in Quine's sense) about the sorts of counterexamples Maher (and Good) have in mind.

[13] Reasonable doubts about the existence (and probative value) of "logical probabilities" traces back (at least) to an earlier debate between Keynes and Ramsey. See [32, §2]. These are thorny issues that we'll have to bracket in this paper. But, see our [5], [14], and [15] for further discussion.

Paradox take a much different tack, and approach The Paradox from a much different perspective. In the next section, we turn to these modern Bayesian approaches.

1.3.2. *Probabilistic Approaches to the Paradox II: Modern, Bayesian Approaches.*

Once probabilists gave up on the Carnapian dream of finding an explication for "logical" conditional probability "$\Pr(H \mid \top)$", they gravitated toward a *subjectivist* approach to probability (and probabilistic confirmation theory). On this view, the conditional probabilities that appear in confirmation-theoretic analyses are *conditional degrees of belief*. That is, "$\Pr(H \mid E)$" now gets interpreted as (something like) *the degree of belief (or degree of confidence) that S assigns to H, on the supposition that E is true.* The agent in question, S, may be an ideally rational agent, in which case the probabilities are supposed to have *pre*scriptive significance. Or, if we're just talking about the conditional degrees of belief that some *actual* agent S happens to have, then the conditional probabilities are more *de*scriptive in nature. In either case, these "subjective" probabilities are much more *psychologistic* than the Carnapian (or Hempelian) confirmation relations we've been talking about so far. As a result, such probabilities are better suited to cognitive-scientific applications such as the Wason Task. In our discussion of the Wason Task, we will say more about the kinds of probabilities that are involved here (especially, their prescriptive *vs* descriptive aspects). But, for now, the important thing is that we do not think of these "subjective" probabilities as *purely* descriptive. Philosophical discussions of the Paradox of Confirmation are, presumably, not offering *mere descriptions* of attitudes actual people happen to have about ravens, *etc.* Rather, they are trying to argue that various attitudes people have to the Paradox either are (or are not) *reasonable* or *rational*. For instance, Hempel and Goodman's "explaining away" is meant to motivate the claim that it would be *reasonable* to *accept* (PC), despite the *apparent* paradoxicality/falsehood of (PC). Moreover, Hempel and Goodman explain why it may *seem reasonable* to reject (PC) by identifying a *closely related* claim (PC*) which it *would* (in fact) be reasonable to *reject*. On the other hand, those who find (PC) paradoxical (or counter-intuitive) may try to argue that the appropriate attitude toward (PC) is that of *rejection*. For instance, Quine thought that (PC) should be *rejected*. As a result, he felt pressure to explain where our simple argument for (PC) *went wrong*. As usual, these philosophers are not engaged in mere description. They are engaged in what one might call *rational reconstruction* (or *rationalization*). This is, of course, a partly *pre*scriptive enterprise.

Contemporary Bayesian approaches to The Paradox are more subtle in their aims. They are not trying to argue for (or rationalize) the acceptability or unacceptability of the *qualitative* confirmation-theoretic claim (PC). As we explained above, that (Hempelian) debate is considered otiose by almost all contemporary Bayesians, because winning it requires a probabilist to countenance "logical" (or, at least, *a priori*) probabilities. But, Bayesians are still interested in rationally reconstructing people's intuitive responses to The Paradox. They just have a different way of doing this. Specifically, Bayesians typically try to motivate a certain *comparative* confirmation-theoretic claim instead. As set-up for a discussion of this comparative approach, we will introduce the following abbreviations: $E_1 \triangleq Ra\&Ba$, $E_2 \triangleq \sim Ba\&\sim Ra$, and $H \triangleq (\forall x)(Rx \supset Bx)$. And, we will use K_α to denote the background corpus of information which consists of our (current) best understanding of the actual world. That is, K_α is allowed to contain whatever we think we know about the actual world. But, we will (for now) assume that K_α does not contain any specific (*de re*) information about the particular object a whose properties are at issue in the evidential claims E_1 and E_2 that will be involved in the comparison.[14] With this set-up in mind, we're now ready to state the comparative claim that most modern Bayesians try to motivate, in their attempt to "rationalize" the Paradox:

($\mathscr{B}$) The degree to which E_2 confirms H relative to K_α is *less than* (perhaps *much* less than) the degree to which E_1 confirms H relative to K_α. Or, more formally, this claim becomes $\mathfrak{c}(H, E_2 \mid K_\alpha) < \mathfrak{c}(H, E_1 \mid K_\alpha)$. And, given our assumption (†), this in turn simplifies to $\Pr(H \mid E_2 \,\&\, K_\alpha) < \Pr(H \mid E_1 \,\&\, K_\alpha)$.

In other words, the Bayesian approach to the Paradox involves trying to establish that ($\mathscr{B}$) $\Pr(H \mid E_2 \,\&\, K_\alpha) < \Pr(H \mid E_1 \,\&\, K_\alpha)$. Let's pause to think about what ($\mathscr{B}$) means for a Bayesian, and why ($\mathscr{B}$) is significant for The Paradox. First, it helps to re-state ($\mathscr{B}$) in English. In English, ($\mathscr{B}$) says that the claim that a is a black raven confers a greater probability upon the claim that all ravens are black than the claim that a is a non-black non-raven does — relative to a background corpus consisting of everything we take ourselves to know about the actual world (excluding specific information about a in particular). Of course, when Bayesians assert ($\mathscr{B}$), they are not merely making some descriptive claim like: "some actual agent S's conditional credences happen to be such that ($\mathscr{B}$) is true of them." Rather, they are making the claim that it would be *reasonable* for an agent S (who resides in the actual world as we know it) to be such that ($\mathscr{B}$) is true of their credences. The idea here is that *if* it is reasonable to have ($\mathscr{B}$)-like credences, *then* this can explain why it is reasonable to think there is something odd (if not paradoxical) about (PC). What's odd about (PC), or so this story might go, is that it (because it is a *coarse-grained, qualitative* claim) obscures the fact that — while E_1 and E_2 might each support H *to some degree* — E_1 constitutes *better evidence* for H than E_2 does. That is to say, while observing a non-black non-raven may be *a* way of generating *some* evidence in favor of H, a *better* way would be to observe a black raven. This more nuanced "explanation" of what *seems* paradoxical about (PC) is not open to Hempel or Goodman, since they have only coarse-grained, qualitative confirmation-theory at their disposal. We won't try (here) to settle the issue of whether establishing ($\mathscr{B}$) would *succeed* in "rationalizing" people's responses to The Paradox (or even in "softening the philosophical impact" of The Paradox). We will just take this for granted here. What we're ultimately interested in here is the probative value of analogous Bayesian "rationalizations" of the responses of (actual) agents to the Wason Task(s). But, first, we need to see *how* Bayesians try to establish ($\mathscr{B}$).

Bayesians have proposed *many* ways of trying to establish ($\mathscr{B}$). But, almost all Bayesian strategies for establishing ($\mathscr{B}$) share the same basic formal structure, which consists of the following three assumptions (or premises) [37].

(3) $\Pr(\sim Ba \mid K_\alpha) > \Pr(Ra \mid K_\alpha)$ $\quad\left[\text{typically, } \frac{\Pr(\sim Ba \mid K_\alpha)}{\Pr(Ra \mid K_\alpha)} \gg 1\right]$

(4) $\Pr(Ra \mid H \,\&\, K_\alpha) = \Pr(Ra \mid K_\alpha)$ $\quad[\therefore \Pr(\sim Ra \mid H \,\&\, K_\alpha) = \Pr(\sim Ra \mid K_\alpha)!]$

(5) $\Pr(\sim Ba \mid H \,\&\, K_\alpha) = \Pr(\sim Ba \mid K_\alpha)$ $\quad[\therefore \Pr(Ba \mid H \,\&\, K_\alpha) = \Pr(Ba \mid K_\alpha)!]$

We will discuss (3)–(5) in some detail, below. But, first, it is important to point out that (3)–(5) are jointly sufficient for the Bayesian comparative claim ($\mathscr{B}$). That is:

[14] For instance, we don't want K_α to include the information that $\sim Ra$ (or Ba) as this might confound the comparison of the confirmatory powers of E_1 *vs* E_2 regarding H. We'll have to relax this restriction on K_α in our discussion of Bayesian approaches to the Wason task, below.

Theorem 1. Claims (3)–(5) jointly entail ($\mathscr{B}$).[15]

In other words, if an agent S's credence function $\Pr(\cdot \mid \cdot)$ satisfies (3)–(5), then it must also (on pain of incoherence) be such that ($\mathscr{B}$) $\Pr(H \mid E_2 \,\&\, K_\alpha) < \Pr(H \mid E_1 \,\&\, K_\alpha)$. That is the *formal* core of the standard Bayesian approaches to The Paradox. Philosophically, the probative value of the Bayesian approach to the Paradox will depend on the status (and philosophical significance) of assumptions (3)–(5).

Before discussing these assumptions, we will present a simple, highly idealized way of *operationalizing* them. Throughout the remainder of the paper, we will adopt a (Good-style) *random-sampling operationalization* of the probability claims concerning Ra, Ba, and H. Here is the idea. We're going to treat the universe (or, at least, those objects in the universe that are accessible to us *via* observation) as a very large urn, and we're going to imagine an agent (S) in a context (C) in which S is about to sample an object a (about which S knows nothing, antecedently) at random from the universe. Now, consider claim (3), for instance. What (3) says is that (given everything S knows about the actual world), it would be reasonable for S to assign a higher probability to the claim that a will turn out to be non-black than the claim that a will turn out to be a raven. Given our random sampling idealization, this makes (3) tantamount to the claim that there is a greater proportion of non-black objects than ravens in the universe (as we know it).

Given this idealization, (3) seems highly plausible. That is, it seems highly plausible that there are (many) more non-black objects than ravens in the (observable) universe. From now on, we will just assume that claim (3) is uncontroversial. Claims (4) and (5), on the other hand, are considerably more controversial. Claim (4) asserts that H (the claim that all ravens are black) is *probabilistically independent* of the claim that a turns out to be a raven. This is equivalent to the claim that H is independent of whether a turns out to be a *non*-raven — a deeply anti-Hempelian claim. Recall that Hempel's dissolution of the Paradox rested on the intuition that $\sim Ra$ is *positively relevant* to H (recall our "indirect support" argument for $\mathscr{E}$). Claim (4) is *incompatible* with this Hempelian intuition. Similarly, Hempel claimed that Ba should be positively relevant to H as well, since Ba also rules-out a as a possible counterexample to H. Thus, claim (5) is also anti-Hempelian in this sense (but, see *fn.* 5). That's not necessarily a reason to reject claims (4) and (5). But, to the extent that you found our ($\mathscr{E}$)-rationale for these Hempelian intuitions plausible, you should also already harbor some doubts about (4) and (5).

At this point, we need to be more careful about the prescriptive *vs* descriptive facets of assumptions (3)–(5). For the Bayesian "rationalization" of the intuitive responses to the Paradox to make sense, it's not enough that some *actual* agent *happen to have* credences that are aligned with (3)–(5). Bayesians also want it to be the case that having credences satisfying (3)–(5) is *epistemically reasonable*. In this context, we recommend thinking of the salient conditional probabilities as *epistemically reasonable* credences — credences that are aligned with what we take to be the operative *epistemological constraints*. So, for instance, if you are inclined to

[15]Owing to limitations of space, we omit proofs of all technical claims made in this paper. We have prepared a *technical addendum*, which explains some of the technical results in detail. The addendum (which includes the proof of a result that vastly generalizes Theorem 1 — see *fn.* 18) can be downloaded from the following URL: `http://fitelson.org/wason_addendum.pdf`.

agree with Hempel that (in the relevant contexts involving the Paradox of Confirmation) learning $\sim Ra$ provides *some support for* H, then this puts pressure on you to *reject* (4). And, there are many other reasons to be worried about (4) and (5).

While it is true that (3)–(5) jointly entail the *comparative* ($\mathscr{B}$), they *also* entail various *qualitative* confirmation-theoretic claims, including the following four:

(6) $\Pr(H \mid Ra \,\&\, Ba \,\&\, K_\alpha) > \Pr(H \mid K_\alpha)$

(7) $\Pr(H \mid \sim Ba \,\&\, \sim Ra \,\&\, K_\alpha) > \Pr(H \mid K_\alpha)$

(8) $\Pr(H \mid Ba \,\&\, \sim Ra \,\&\, K_\alpha) < \Pr(H \mid K_\alpha)$

(9) $\Pr(H \mid \sim Ba \,\&\, \sim Ra \,\&\, K_\alpha) > \Pr(H \mid \sim Ra \,\&\, K_\alpha)$

Claim (6) says that $Ra \,\&\, Ba$ confirms that all ravens are black. Since $Ra \,\&\, Ba$ is a "positive instance" of the claim that all ravens are black, (6) constitutes the "normal" or "direct" application of (NC) to the hypothesis that all ravens are black. Claim (7) is just the "paradoxical conclusion" (PC) — that $\sim Ba \,\&\, \sim Ra$ confirms that all ravens are black. In other words, assumptions (3)–(5) entail that an agent S *must* accept that *both* $Ra \,\&\, Ba$ *and* $\sim Ba \,\&\, \sim Ra$ lend *some* support to H. This is a rather Hempelian consequence. And, it may be an acceptable consequence (if, for instance, our "indirect support" rationale for $\mathscr{E}$ is truly compelling). But, it is somewhat odd that assumptions which were designed to ensure a *comparative* claim ($\mathscr{B}$) also commit the Bayesian to traditional *qualitative* claims that are very much in the Hempelian spirit of (NC) and (PC). As Good and Maher have taught us, such qualitative commitments are *not* — in general — *forced* on a Bayesian. But, if one buys into the standard Bayesian line on the Paradox, then one *is* saddled with these very commitments. This suggests that (4) and (5) are rather strong. And, when one considers (8) and (9), one sees that (4) and (5) are *implausibly* strong.

What does claim (8) say? It says that $Ba \,\&\, \sim Ra$ ***dis***confirms the hypothesis that all ravens are black. That is quite a puzzling consequence of (3)–(5). After all, if we learn that a is a black non-raven, then we are learning (twice over!) that a cannot be a counterexample to H. So, the "indirect support" ($\mathscr{E}$)-rationale that underlies the Hempelian commitments to (6) and (7) [*viz.*, to (NC) and (PC)] *doubly cuts against* a commitment to (8).[16] Moreover, independently of Hempelian considerations, it seems very odd that a Bayesian who is interested in establishing a *comparative* claim such as ($\mathscr{B}$) should find themselves committed a *qualitative* claim that sounds as strange as (8) does. This reveals just how strong the independence assumptions (4) and (5) really are. Finally, consider consequence (9). This is just the Bayesian analogue of Hempel's (PC*). Recall that Hempel's intuition was that (PC*) should come out *false*. In Bayesian terms, that Hempelian intuition is:

($\mathscr{H}$) $\Pr(H \mid \sim Ba \,\&\, \sim Ra \,\&\, K_\alpha) \leq \Pr(H \mid \sim Ra \,\&\, K_\alpha)$

The reason Hempel thought ($\mathscr{H}$) should be true [and therefore claim (9) should be *false*] is that it is $\sim Ra$ that provides the (indirect) "confirmational boost" to H. So, if the agent *already knows* $\sim Ra$, then learning (in addition) that $\sim Ba$ should not provide any *further* support for H. In Bayesian terms, this boils down to ($\mathscr{H}$). Unfortunately, the standard Bayesian assumptions (3)–(5) jointly entail (9), which is *incompatible* with Hempel's intuition ($\mathscr{H}$); and, more generally, with Hempel's

[16]It seems that the first person to explicitly *deny* (8) was Keynes [21, *p*. 230]. In general, Keynes does not get enough credit in this context as he should. Many of the main principles of (early) inductive logic were first discussed by Keynes. But, subsequent work of Nicod [28] and Hempel [17] on confirmation theory seems to have had the effect of screening us off from Keynes's original work.

claim that (PC) is true while (PC*) is false. Whether we agree or disagree with these tenets that are central to the Hempel/Goodman "explaining away" of the Paradox, it is certainly (at the very least) strange that the Bayesian *comparative* strategy should saddle us with yet another historically controversial *qualitative* claim.

To sum up: (4) and (5) are *sufficient* [in conjunction with the highly plausible (3)] to ensure the desired comparative claim ($\mathscr{B}$), but they also saddle the Bayesian with undesirable *qualitative* confirmation-theoretic commitments.[17] It would be nice if there were a way to *weaken* assumptions (4) and (5), so as to avoid these undesirable qualitative conclusions, while preserving the desired *comparative* conclusion ($\mathscr{B}$). As we have recently discovered [8], there is a way to do just that.

1.3.3. *Probabilistic Approaches to the Paradox II: A New Bayesian Approach.*

As we have seen, the standard Bayesian approach to the paradox involves the identification of constraints on an agent S's credence function that suffice to ensure that the comparative claim ($\mathscr{B}$) is true, thus ensuring that $Ra \,\&\, Ba$ constitutes *better evidence* (for S) for the claim that all ravens are black than $\sim Ba \,\&\, \sim Ra$ does. Unfortunately, as we have also seen, the standard constraints (3)–(5) that are used in traditional Bayesian comparative approaches to the paradox have various undesirable qualitative consequences (*e.g.*, that $Ba \,\&\, \sim Ra$ ***dis***confirms the hypothesis that all ravens are black). We have recently discovered [8] that the controversial independence assumptions (4) and (5) can be *significantly weakened*, so as to avoid all of their undesirable qualitative consequences, while preserving the desired comparative conclusion ($\mathscr{B}$). To wit, consider the following constraint:

$$(\ddagger)\quad \frac{\Pr(\sim Ba \mid K_\alpha)}{\Pr(Ra \mid K_\alpha)} \gtrapprox \frac{\Pr(\sim Ba \mid H \,\&\, K_\alpha)}{\Pr(Ra \mid H \,\&\, K_\alpha)}$$

Where the relation "$\gtrapprox$" has the following meaning:

$$(\ddagger)\quad \textit{Either } \frac{\Pr(\sim Ba \mid K_\alpha)}{\Pr(Ra \mid K_\alpha)} \geq \frac{\Pr(\sim Ba \mid H \,\&\, K_\alpha)}{\Pr(Ra \mid H \,\&\, K_\alpha)},$$

or

$$\frac{\Pr(\sim Ba \mid H \,\&\, K_\alpha)}{\Pr(Ra \mid H \,\&\, K_\alpha)} \text{ is } \textit{slightly}[18] \text{ larger than } \frac{\Pr(\sim Ba \mid K_\alpha)}{\Pr(Ra \mid K_\alpha)}.$$

What (‡) says is that *learning H does not dramatically increase one's estimate of the ratio of non-black objects to ravens in the universe.* Typically, it is assumed [(3)] that the ratio of non-black objects to ravens in the universe is large [*viz.*, $\frac{\Pr(\sim Ba|K_\alpha)}{\Pr(Ra|K_\alpha)} \gg 1$]. What (‡) says is that our estimate of this ratio should not be made substantially larger (merely) by learning that all ravens are black. Now, a few remarks about (‡).

First, note that (‡) is *strictly weaker than* the conjunction of (4) and (5), because (4) & (5) entails $\Pr(\sim Ba \mid K_\alpha) = \Pr(\sim Ba \mid H \,\&\, K_\alpha)$ and $\Pr(Ra \mid K_\alpha) = \Pr(Ra \mid H \,\&\, K_\alpha)$.

[17] See Peter Vranas's excellent paper on The Paradox [37] for an in-depth discussion of the independence assumptions (4) and (5), and various other ways in which they seem epistemically implausible.

[18] The closeness of approximation that is required in (‡) will depend on *how large* the ratio $\frac{\Pr(\sim Ba|K_\alpha)}{\Pr(Ra|K_\alpha)}$ in condition (3) is. If this ratio is *very large*, then the "$\approx$" in (‡) needn't be as close as if this ratio is only slightly greater than 1. Our initial technical results concerning ($\mathscr{B}$) — which are couched in terms of comparisons of posterior probabilities (as opposed to likelihood ratios) and which involve the stronger relation "$\geq$" — are reported in [8]. We have since identified *necessary and sufficient conditions* for ($\mathscr{B}$), which subsume all results in this area as special cases (including the present results). See the *technical addendum* for the formal details ⟨`http://fitelson.org/wason_addendum.pdf`⟩.

Claim (‡), on the other hand, does *not* imply any independence relations between H, Ra, and/or Ba, nor does (‡) even entail that the *ratios* $\frac{\Pr(\sim Ba|K_\alpha)}{\Pr(Ra|K_\alpha)}$ and $\frac{\Pr(\sim Ba|H\&K_\alpha)}{\Pr(Ra|H\&K_\alpha)}$ are *exactly* equal. Second, note that (‡) has the following two crucial properties:

- (3) and (‡) *entail* ($\mathscr{B}$). To be more precise, given (3), ($\mathscr{B}$) is true whenever $\frac{\Pr(\sim Ba|K_\alpha)}{\Pr(Ra|K_\alpha)} \geq \frac{\Pr(\sim Ba|H\&K_\alpha)}{\Pr(Ra|H\&K_\alpha)}$. Moreover, given (3), ($\mathscr{B}$) *remains true even if* $\frac{\Pr(\sim Ba|H\&K_\alpha)}{\Pr(Ra|H\&K_\alpha)}$ is *slightly larger than* $\frac{\Pr(\sim Ba|K_\alpha)}{\Pr(Ra|K_\alpha)}$ — where we state *precisely* what "slightly" means in the *technical addendum* to this paper (see *fn.* 18).
- (3) and (‡) do *not* jointly entail *any* of the qualitative claims (6)–(9). More precisely, the conjunction (3) & (‡) is compatible with *either asserting or denying* any of the four qualitative claims (6)–(9).

In other words, in conjunction with (3), (‡) *lacks* the *un*desirable consequences that (4) & (5) have. This suggests that (‡) is *more* reasonable than (4) & (5) as an additional constraint on credences, for the purposes of providing a comparative "rationalization" or "softening" of The Paradox of Confirmation. We won't try to argue here that (‡) is reasonable *simpliciter* [presently, we will be content with the *comparative* claim that (‡) is *more plausible* than the conjunction (4) & (5)]. But, we will return to (‡), and related assumptions, below, when we discuss analogous Bayesian "rationalizations" of the responses of subjects faced with the Wason Task(s).

2. The Wason Task(s)

Here is one of Wason's original descriptions of his reasoning task [38]:

> Given the sentence: Every card which has a D on one side has a 3 on the other side (and knowledge that each card has a letter on one side and a number on the other), together with four cards showing D, K, 3, 7, hardly any individuals make the correct choice of cards to turn over (D, 7) in order to determine the truth of the sentence.

The first thing we'll need to do is make our working description of this task *more precise.* Once we've clarified the task(s) implicit in Wason-style experiments, we will look at some empirical results involving the responses of actual agents to various clarified versions of the task(s). Then, we will examine some analyses of these responses. We'll be most interested in Bayesian "rationalizations" of the responses, but we'll also discuss some other analyses along the way. Our main goal will be to bring out the analogies (and disanalogies) between Wason's Task(s) and The Paradox of confirmation. Step one: clarification of the Wason Task(s).

2.1. **Clarifying Wason's Task(s).**

Here is a clearer characterization of the task that Wason (sketchily) described:

> Each card (in some set of cards **C**) has one letter on one side and one number on the other side. You will be shown four cards from **C** (with one face down), and you will be asked to turn over one or more of the four cards, with an eye toward determining whether the following hypothesis is true:
>
> (*H*) All "*D*"-cards (in **C**) are "3"-cards.
>
> *Q*: Which of the following 4 cards would you turn over, in order to test *H*?
>
> | *D* | *K* | 3 | 7 |
> |---|---|---|---|

This is a pretty good "first pass" at a clarified Wason Task. But, it still needs work.

First, as Humberstone [20] points out, there is still an important ambiguity in this statement of the task, having to do with **C** — *the domain of quantification* of H. Humberstone distinguishes the following two interpretations of **C**, which are (intuitively) both compatible with Wason's original description, and which also cohere (although perhaps to a lesser extent) even with our clarified description:

(I) $\mathbf{C} = \{\boxed{D}\ \boxed{K}\ \boxed{3}\ \boxed{7}\}$.

(II) $\{\boxed{D}\ \boxed{K}\ \boxed{3}\ \boxed{7}\} \subset \mathbf{C}$.

On interpretation (I), the domain of quantification **C** of H is *identical to* the set of four cards that the subject is shown. This seems to be what *Wason* has in mind, since, on this reading, it is clear that there is a *uniquely best* (most probative) choice of cards to turn over for the purpose of testing H: namely, $\boxed{D}$ and $\boxed{7}$. Why? Well, if the four cards you are shown are *all the cards there are*, then *the only possible counterexamples to* H are $\boxed{D}$ and $\boxed{7}$. So, by turning over those two cards (and learning what's on their other sides), you are *guaranteed* to *definitively establish* the truth-value of H. Wason speaks of "the" correct choice and "determining" the truth of H. This suggests that he had in mind Humberstone's interpretation (I).

On the other hand, interpretation (II) leads to a much less definitive task. If the four cards you are shown are a *proper subset* of the set of all cards that exist, then there is no longer *any* choice of cards to turn over that will *guarantee* a *decisive* test of H. Moreover, it seems plausible that many subjects will be more likely to hear the task as being of type (II), as opposed to type (I). After all, our experiences with "cards" usually involve decks which contain more than four cards, and which have been "shuffled" so as to induce a kind of *random sampling* behavior.

Because our aim is to forge as good an analogy with the Paradox of Confirmation as possible, we will assume Humberstone's interpretation (II). In addition to removing Humbertone's ambiguity, it will also be useful to clarify the talk of "testing" and "determination of truth" in our descriptions. Combining these refinements yields:

> Each card (in some deck of cards **C**, which contains many cards) has one letter on one side and one number on the other side. The deck has been well-shuffled, and four cards have been drawn from the deck and placed on the table in front of you (with one side visible). Specifically, you see:
>
> $\boxed{D}\ \boxed{K}\ \boxed{3}\ \boxed{7}$
>
> Now, consider the following hypothesis about the cards in the deck **C**:
>
> (H) All "D"-cards (in **C**) are "3"-cards.
>
> Your task is to turn over one ore more of these four cards, with any eye toward testing H in the most effective and efficient way you can. That is, you want to turn over those cards which you think are most evidentially relevant to the hypothesis H (and none that are irrelevant to H).

This description of the task is more precise, and it also allows for the tightest analogy with the Paradox of Confirmation. We will discuss some minor variations, but this will be the basic sort of description we will use when we talk about Wason.

2.2. **The Analogy Between The Wason Task and The Paradox of Confirmation.**

Astute readers may already be able to anticipate where this "analogy" is going. But, there will be a few twists and turns along the way. The way we are going to forge

the analogy is by re-describing the Paradox of Confirmation, so as to turn *it* into a Wason Task, in the above sense. Here's our re-description of The Paradox:

> For each object in the world, we will create a card. On one side of the card, we will write "R" (or "$\sim R$"), depending on whether the object is (or is not) a raven. On the other side of the card, we will write "B" (or "$\sim B$"), depending on whether the object is (or is not) black. Then, we will take the resulting deck of cards (**C**) and make sure it is well shuffled. Finally, we will draw four cards from this deck and place them on a table in front of you (with one side visible). And, you end-up seeing the following:
>
> $\boxed{R}\ \boxed{\sim R}\ \boxed{B}\ \boxed{\sim B}$
>
> Now, consider the following hypothesis about the cards in the deck **C**:
>
> (H) All "R"-cards (in **C**) are "B"-cards. (*i.e.*, all ravens are black.)
>
> Your task is to turn over one ore more of these four cards, with any eye toward testing H in the most effective and efficient way you can. That is, you want to turn over those cards which you think are most evidentially relevant to the hypothesis H (and none that are irrelevant to H).

As you can see, this version of The Paradox is *perfectly* analogous with The Wason Task (as we have clarified it above). But, this version of The Paradox differs from the original Paradox of Confirmation in the following two crucial ways.

- In this version, the subjects are allowed to turn over *multiple* cards. But, in the original version of The Paradox, only *one* object is randomly sampled.
- In this version, the subject already knows some (salient) properties of the objects that have been sampled. But, in the original version, this sort of specific information about the sampled object was *not* included in K_α.

We're just going to have to live with the second disanalogy. But, the Bayesian apparatus we have already been employing will have no problem coping with additional information (*i.e.*, specific information about the sampled card a in particular) appearing in the background. This will be handled in the usual way, by *conditionalizing* on the additional information.[19] We have already seen an example of this in claim (9), above, which (essentially) adds $\sim Ra$ to the background corpus. This is rather non-Hempelian, but we'll just have to leave this Hempelian caveat behind.

We will shore-up the first disanalogy by focusing on *single*-card-turning strategies. That is, we will only be comparing strategies that turn over *exactly one* of the four visible cards. This is a limitation that could be relaxed, but it would make the analyses much more complicated. In any event, this restriction to single-card strategies is a standard move that is made in the (Bayesian) cognitive science literature on the Wason Task. Next, we'll look at some *empirical Wason data*.

2.3. **Empirical Data Concerning the Wason Task(s).**

Over the past 40 years or so, various versions of Wason's Task(s) have been given to many actual subjects around the world. The data are pretty robust, across the different variations of the Task(s) we've been considering. Whether it be Wason's original descriptions, or descriptions much closer to the one we're adopting here,

[19] In fact, some Bayesian approaches to The Paradox of Confirmation *already* have this "two-stage sampling" structure. See, for instance, [3, *pp*. 69–73] and [33, appendix].

the patterns of response are usually the same.[20] In decreasing order of frequency, the responses to the Wason Task are as follows, with the analogous responses for the R/B version of the Wason Task in brackets.[21]

(*i*) $\boxed{D}\ \boxed{3}$. $\left[\boxed{R}\ \boxed{B}\right]$

(*ii*) $\boxed{D}$. $\left[\boxed{R}\right]$

(*iii*) $\boxed{D}\ \boxed{3}\ \boxed{7}$. $\left[\boxed{R}\ \boxed{B}\ \boxed{\sim B}\right]$

(*iv*) $\boxed{D}\ \boxed{7}$. $\left[\boxed{R}\ \boxed{\sim B}\right]$

Note that Wason's "correct answer" $\boxed{D}\ \boxed{7}\ \left[\boxed{R}\ \boxed{\sim B}\right]$ is at the *bottom* of this list.

Since we will be focusing on *single*-card-turning strategies, we'll need some empirical data about the actual frequency of responses, in terms of *single*-card-turning strategies. There are various ways to obtain such data. We could do a meta-analysis over the standard (multiple-card) Wason experiments, and try to extract single-card-strategy ranking data. Alternatively, we could look at other variants of the Task, in which subjects are explicitly asked to rank single card strategies, in terms of their relative evidential relevance regarding (or probative value for testing) H. It doesn't matter very much which way we do this. The empirical ranking of single-card strategies is pretty robust across different (reasonable) ways of obtaining such data.[22] Here is the empirical ranking for single-card strategies (again, with the analogous empirical ranking for the R/B versions of the Task in brackets).

(*i*) $\boxed{D}$. $\left[\boxed{R}\right]$

(*ii*) $\boxed{3}$. $\left[\boxed{B}\right]$

(*iii*) $\boxed{7}$. $\left[\boxed{\sim B}\right]$

You probably could have guessed this on the basis of "top–4" of the full ranking of all strategies, reported above. In any case, we will be focusing just on the "top-3" of the full ranking of single-card strategies (we won't concern ourselves with the last place single-card strategy $\boxed{\sim R}$, since that's not where the controversies lie here). That is, we will focus on the following ordering of single-card strategies:

($\mathbb{O}$) $\boxed{R} \succ \boxed{B} \succ \boxed{\sim B}$

The aim of Bayesian approaches to The Wason Task will be to "rationalize" or "explain" this preference ordering ($\mathbb{O}$) among single-card strategies exhibited by actual subjects. From now on, we're going to talk about the R/B version of the task, rather than the $D/3$ version of the task [25]. This will allow us to use the same notation and results we've already developed in our discussion of The Paradox.

[20] If, instead of talking about abstract, extensional hypotheses involving numbers, letters, birds, colors, *etc*., we instead offer subjects Wason Task(s) involving hypotheses with *deontic* or *modal* content, then there is evidence that this can significantly alter the patterns of response to the Wason Task [2]. We won't have the space to discuss such "content effects" here. But, it is worth noting that the possibility of similar "content effects" has been raised in connection with The Paradox of Confirmation as well [36]. We will restrict our attention to the case of non-modal, classical, extensional generalizations involving ravens and blackness. In such renditions of the Wason Task, the patterns of response discussed here are pretty robust across our variations in description.

[21] See [25] for experiments performed explicitly on R/B versions of the Task.

[22] See Oaksford and Chater's [29] for a discussion of obtaining single-card strategy ranking data.

2.4. **Bayesian Approaches to The Wason Task I: Mind the Gap.**

In this section, we will use what we already know from our discussion of Bayesian approaches to The Paradox to try to formulate an analogous analysis of The Wason Task. As we'll see, what we already know is not quite enough to provide an adequate "rationalization" (or even a "how-possibly explanation") of the empirical ordering ($\mathbb{O}$) of single-card strategies in The Wason Task. There is a *gap* between what we can do with the Bayesian analyses of The Paradox, and what we need to be able to do in order to "rationalize" or "explain" ($\mathbb{O}$). In this section, we'll explain what that gap is. In the next section, we'll discuss a recent Bayesian approach to the The Wason Task which fills this gap with some additional Bayesian technology.

The obvious question at this stage is: Do either of the Bayesian approaches to The Paradox that we discussed above have the wherewithal to account for ($\mathbb{O}$)? That is, do either (3)–(5) or (3) & (‡) *entail* a *confirmational* ordering that is in agreement with the *empirical* ordering ($\mathbb{O}$)? Before we answer this question, we need to get clearer on what it would *mean* for a confirmational ordering to "agree with" ($\mathbb{O}$).

Let's start out by being a little less ambitious. Let's focus just on the following *fragment* of the empirical ordering ($\mathbb{O}$): $\boxed{R} \succ \boxed{\sim B}$. What $\boxed{R} \succ \boxed{\sim B}$ says is that turning over the R-card generates (in some sense) "better evidence" regarding H than turning over the $\sim B$ card does. The scare quotes on "better evidence" are there to indicate that — in the present context — "better evidence" can*not* just mean "evidence *which confirms H more strongly*". Why not? Because single-card strategies are capable of generating *qualitatively different sorts* of evidence, depending on their outcomes. For instance, suppose you turn over the R-card (*i.e*., you choose the $\boxed{R}$ strategy for testing H). There are two possibilities. The other side could say "$\sim B$", in which case H will have been *refuted*, or, the other side could say "B", in which case you come to learn that the card a is a "positive instance" of H. Similarly, if you turn over the $\sim B$-card, then you could get an "R", which would *refute* H, or, you could get a "$\sim R$", which would be tantamount to learning that a is (a card which corresponds to) a non-black, non-raven. Thus, when comparing the $\boxed{R}$ and $\boxed{\sim B}$ strategies using the methods we've been employing with the Paradox, the best we can do is compare *outcomes* of applying those strategies.

To denote these outcomes, we will add superscripts to the boxes. So, for instance, we will write $\boxed{R}^{\sim B}$ to denote the outcome of applying the $\boxed{R}$ strategy *and then discovering that the other side of the R-card says "$\sim B$"*. With this notation in mind, we can now (meaningfully) talk about the (traditional) confirmation-theoretic orderings for each of the pairs of possible outcomes of the two strategies $\boxed{R}$ and $\boxed{\sim B}$. *Some* pairs of outcomes involving $\boxed{R}$ *vs* $\boxed{\sim B}$ have a clear comparative confirmational structure. For instance, the following two claims should come out true:

- $\boxed{R}^{B}$ confirms H more strongly than $\boxed{\sim B}^{R}$ does.
- $\boxed{\sim B}^{\sim R}$ confirms H more strongly than $\boxed{R}^{\sim B}$ does.

Intuitively, the first claim should come out true because $\boxed{\sim B}^{R}$ *refutes* H, but $\boxed{R}^{B}$ doesn't. Similarly, the second claim should come out true because $\boxed{R}^{\sim B}$ *refutes* H, but $\boxed{\sim B}^{\sim R}$ doesn't. But, what can we say about $\boxed{R}$ *vs* $\boxed{\sim B}$ *simpliciter*? Not much. This is *the gap* between what the Bayesian methods we applied to The Paradox *can* do, and what a proper Bayesian analysis of The Wason Task *needs* to

be able to do. In the next section, we will discuss a technique for "confirmation-theoretically averaging" over possible outcomes of single-card strategies, so as to obtain meaningful *overall* comparisons of *entire* strategies. This will *fill the gap*.

2.5. **Bayesian Approaches to The Wason Task II: Nickerson's Filling of the Gap.**

In a recent cognitive science paper on the Wason Task, Nickerson [26] provides a Bayesian way of filling the gap we pointed out in the last section. Nickerson's key idea is to define a notion of *expected (or average) confirmational power* for single-card strategies. The first step is to define a measure of *confirmational power*. Specifically, Nickerson adopts the following measure $\bar{d}(H, E \mid K)$ of *the confirmational power of E regarding H, relative to background corpus K*:

$$\bar{d}(H, E \mid K) \triangleq |\Pr(H \mid E \,\&\, K) - \Pr(H \mid K)|$$

Nickerson's $\bar{d}(H, E \mid K)$ is just the absolute value of what is known as the difference measure $d(H, E \mid K)$ of the degree to which E *confirms* H, relative to K. If $d(H, E \mid K)$ is positive, then E confirms H, relative to K (in the standard, Bayesian sense). If $d(H, E \mid K)$ is negative, then E *dis*confirms H, relative to K. And, if $d(H, E \mid K) = 0$, then E is said to be *confirmationally neutral* (or *irrelevant*) to H, relative K. So, by taking the *absolute value* of d, Nickerson is defining a measure of *confirmational power*, where "power" is *neutral* with respect to whether we're talking about power to confirm or power to disconfirm. Another way of thinking about confirmational power is to think of it as a measure of how *evidentially relevant* E is to H (relative to K), where this means *either positively or negatively* evidentially relevant. Confirmational power is exactly the sort of quantity we need to fill the gap from the last section. Because strategies like $\boxed{R}$ can generate either positively or negatively relevant evidence regarding H (depending on which *outcome* of $\boxed{R}$ obtains), we can now compare both types of evidence on a *single scale* of confirmational power.

We need just one more ingredient, and that is a precise Bayesian definition of *the expected (or average) confirmational power of a strategy s*: $\mathcal{P}(s)$. Bayesian statistical decision theory has a standard way of defining $\mathcal{P}(s)$, and Nickerson follows this traditional recipe.[23] For instance, here is Nickerson's definition of $\mathcal{P}\big(\boxed{R}\big)$.[24]

$$\mathcal{P}\big(\boxed{R}\big) \triangleq \Pr(Ba \mid Ra) \cdot \bar{d}(H, Ba \mid Ra) + \Pr(\sim Ba \mid Ra) \cdot \bar{d}(H, \sim Ba \mid Ra)$$

Let's think about why this makes sense as a definition of $\mathcal{P}\big(\boxed{R}\big)$. As we saw in the last section, there are two possible outcomes of an application of the $\boxed{R}$ strategy.

(1) $\boxed{R}^{B}$. Here, Nickerson's degree of confirmational power is $\bar{d}(H, Ba \mid Ra)$, since $d(H, Ba \mid Ra)$ is the degree to which Ba confirms H, relative to background corpus Ra (according to Nickerson's difference measure). The reason Ra is in the background corpus here is that we *already know* that Ra is true, since we can see that side of card a before we turn the card over.

[23] I.J. Good and Alan Turing were among the first Bayesian statisticians to employ this sort of "expected confirmational power" concept. Specifically, Good & Turing made use of the *expected weight of evidence* of experiments in their work on the Enigma project in WWII. While Good & Turing used a different underlying measure of degree of confirmation than Nickerson's d (specifically, they used the *log-likelihood-ratio* measure), their basic idea was the same. See [11] for discussion.

[24] Hereafter, we will often suppress the "K_α"s from the antecedents of conditional probabilities and confirmation functions. It is to be understood that all Pr's/c's are (implicitly) *conditionalized on* K_α.

(2) $\boxed{R}^{\sim B}$. Here, Nickerson's degree of confirmational power is $\bar{d}(H, \sim Ba \mid Ra)$, since $d(H, \sim Ba \mid Ra)$ is the degree to which $\sim Ba$ confirms H, relative to background corpus Ra (according to Nickerson's difference measure).

Then, $\mathcal{P}\big(\boxed{R}\big)$ is just a *weighted average* of these two possible values of confirmational power, where the weights are the appropriate conditional probabilities. In case (1), we weight the confirmational power $\bar{d}(H, Ba \mid Ra)$ by $\Pr(Ba \mid Ra)$, since that's the probability of obtaining a B-outcome in an application of the $\boxed{R}$ strategy. Similarly, in case (2), we weight the confirmational power $\bar{d}(H, \sim Ba \mid Ra)$ by $\Pr(\sim Ba \mid Ra)$, since that's the probability of obtaining a $\sim B$-outcome in an application of the $\boxed{R}$ strategy. Similar definitions are used for the other two strategies:

$$\mathcal{P}\big(\boxed{B}\big) \triangleq \Pr(Ra \mid Ba) \cdot \bar{d}(H, Ra \mid Ba) + \Pr(\sim Ra \mid Ba) \cdot \bar{d}(H, \sim Ra \mid Ba).$$

$$\mathcal{P}\big(\boxed{\sim B}\big) \triangleq \Pr(Ra \mid \sim Ba) \cdot \bar{d}(H, Ra \mid \sim Ba) + \Pr(\sim Ra \mid \sim Ba) \cdot \bar{d}(H, \sim Ra \mid \sim Ba).$$

With these definitions in hand, we are now (finally) ready to say what it *means* for a *Nickersonian* confirmational ordering to "agree with" ordering ($\mathbb{O}$). It's just what you would expect. Recall, the empirical ordering of strategies is $\boxed{R} \succ \boxed{B} \succ \boxed{\sim B}$. Thus, a Nickersonian confirmational ordering will agree with ($\mathbb{O}$), just in case:

($\mathcal{N}$) $\mathcal{P}\big(\boxed{R}\big) > \mathcal{P}\big(\boxed{B}\big) > \mathcal{P}\big(\boxed{\sim B}\big)$.

So, one thing Nickerson needs to do, in order to complete his Bayesian analysis of the Wason task, is to identify constraints on an agent's credence function *which jointly entail* ($\mathcal{N}$). And, Nickerson does just that. Specifically, Nickerson explicitly models his analysis on the traditional Bayesian approach to The Paradox of Confirmation. That is, he takes *as his starting point*, assumptions (3)–(5), above. Unfortunately, as we will explain below, (3)–(5) alone *do not entail* ($\mathcal{N}$). So, Nickerson adds-in additional numerical constraints, so as to end-up with a *specific probability model* on which ($\mathcal{N}$) comes out true. We will not get into the numerical details of Nickerson's specific probability model. But, in the next section, we will offer some criticisms of (and friendly amendments to) Nickerson's approach, which draw on the insights gleaned from our historical study of The Paradox of Confirmation.

2.6. **Bayesian Approaches to The Wason Task III: Critique of Nickerson.**

We think Nickerson has the right *kind* of Bayesian-confirmation-theoretic approach to the Wason Task. And, we also think he is right to look to the historical literature on the Paradox of Confirmation for inspiration. But, from our perspective, his *specific implementation* of this kind of approach has several shortcomings.

First, Nickerson endorses the traditional Bayesian approach to The Paradox, which is committed to the very strong assumptions (3)–(5). We have argued that these assumptions are (from an epistemic point of view) *implausibly* strong. But, does this imply that Nickerson's use of (3)–(5) is inappropriate? This may depend on what Nickerson is using (3)–(5) *for*. Is Nickerson trying to "rationalize" the responses of actual subjects, or is he merely trying to offer a "how possibly explanation" — that is, trying to explain how a Bayesian *can coherently* have a credence function that satisfies his confirmational power ordering ($\mathcal{N}$)? If he is merely doing the latter, then taking (3)–(5) *as a starting point* for constructing ***a*** coherent set of credences that satisfy ($\mathcal{N}$) may be acceptable. Moreover, even if he means to be "rationalizing" the responses, he could always retreat to a radical subjectivist

Bayesian line, according to which the *only* prescriptive constraints an agent's credences are *the probability axioms*. We are not sympathetic with this sort of radical subjective Bayesianism. And, as we have argued above, we think that (3)–(5) have some *epistemically unreasonable* consequences. But, perhaps Nickerson's story is not meant to have *any pre*scriptive content. Perhaps it *is merely* a "how possibly" story. If that's right, then we think this is *itself* a shortcoming. After all, most Bayesians who talk about the Paradox of Confirmation believe that certain attitudes toward the paradox would be *unreasonable*, even though they don't entail violations of probabilistic *coherence*. And, presumably, Wason himself was interested in the behavior of his subjects *because* he thought they tended to give the *wrong* answer to his question. So, it seems to me that an approach to the Wason Task (especially one which borrows heavily from the philosophical literature on The Paradox of Confirmation) should *not* be devoid of prescriptive content. Thus, Nickerson faces a dilemma. *Either* he is doing something which is not that interesting (telling a *mere* "how possibly story"), *or* he is doing something interesting, but in a prescriptively implausible way. We won't try to settle this question here.[25]

There are other problems with Nickerson's approach. His discussion is not very illuminating when it comes to identifying *general* conditions under which ($\mathscr{N}$) obtains. Basically, he writes down a numerical probability model — inspired by the traditional Bayesian assumptions (3)–(5) — that happens to satisfy ($\mathscr{N}$), without really explaining why the model satisfies ($\mathscr{N}$). To be more specific, Nickerson seems to be aware that (3)–(5) *alone* do *not* suffice for ($\mathscr{N}$), but he doesn't offer any analysis of what needs to be super-added to (3)–(5) to get the job done. As it turns out, there is a very simple explanation. All we need to do is strengthen (3) to:

(3′) $\Pr(\sim Ba \mid K_\alpha) > \Pr(Ba \mid K_\alpha) > \Pr(Ra \mid K_\alpha)$.

Given our random sampling setup, all (3′) says is that there are more non-black things in the universe than there are black things, and there are more black things than there are ravens. This, of course, entails (3). And, more interestingly, we have:

Theorem 2. (3′)–(5) jointly entail ($\mathscr{N}$).

So, this is one very simple way of augmenting the standard Bayesian story about The Paradox, so as to yield the desired Nickersonian confirmational power ordering ($\mathscr{N}$). Of course, we don't think this approach is epistemically plausible, since it still has all of the epistemically undesirable consequences of the standard Bayesian approach. But, it's a very simple, general way of ensuring ($\mathscr{N}$), which requires only a very minor tweak to the old Bayesian approaches to the Paradox of Confirmation.

What about our new-and-improved Bayesian approach to the Paradox, which avoids the epistemically undesirable consequences of the standard approach? Is there an (exactly) analogous weakening of Nickerson's assumptions, which would avoid the undesirable consequences of (3′)–(5), while preserving Nickerson's desired conclusion ($\mathscr{N}$)? Unfortunately, the answer is *no*, because (3′) and (‡) do *not* jointly entail ($\mathscr{N}$). However, (3′) and (‡) do entail *part* of ($\mathscr{N}$). To wit:

Theorem 3. (3′) and (‡) jointly entail that $\mathcal{P}\big(\boxed{R}\big) > \mathcal{P}\big(\boxed{\sim B}\big)$.

[25]Oaksford and Chater [29, 30] use similar Bayesian techniques (and some assumptions similar to assumptions made in traditional Bayesian approaches to The Paradox), explicitly with an eye toward showing that the empirical ordering (ℚ) of Wason responses can be seen as *rational*. Nickerson is less clear about his prescriptive commitments, but his approach must have *some* prescriptive significance. We have chosen to focus on Nickerson's approach here, because it is the *most* continuous with (and sensitive to) the philosophical and historical literature on the Paradox of Confirmation.

In other words, (3′) and (‡) are sufficient to guarantee that turning over the $\boxed{R}$-card will provide *better evidence on average* than turning over the $\boxed{\sim B}$-card will. This is a nice, two-stage-sampling rendition of our new-and-improved Bayesian account of the Raven Paradox. But, unfortunately, (3′) and (‡) are too weak to entail anything about the place of $\boxed{B}$ in the $\mathcal{P}$-ordering. And, *that's* where the controversy lies.

This raises an interesting and important theoretical question. What is the weakest condition (C) such that (3′) & (‡) & (C) *does* entail ($\mathscr{N}$)? While we don't have a (complete) answer to that question (yet), we have been able to identify a very interesting *necessary condition* for ($\mathscr{N}$) — *given* (3′) and (‡), *plus* the following (highly plausible) additional (prescriptive) confirmation-theoretic assumption:

($\star$) $\bar{\mathfrak{c}}(H, Ra \mid \sim Ba) > \bar{\mathfrak{c}}(H, Ra \mid Ba)$.

In our box-notation, ($\star$) says that $\boxed{\sim B}^R$ should have greater *confirmational power* (regarding H) than $\boxed{B}^R$ does. This seems highly plausible, since $\boxed{\sim B}^R$ *refutes* H, while $\boxed{B}^R$ is a mere "positive instance" of H. The basic idea here is just that *refuting instances are more powerful than positive instances.*[26] Now, a crucial result:

Theorem 4. *Given* (3′), (‡), *and* ($\star$), the following is *necessary* for ($\mathscr{N}$):

$$\Pr(Ra \mid Ba) > \Pr(Ra \mid \sim Ba)$$

In other words, taking the three conditions (3′), (‡), and ($\star$) as background assumptions, *the only way* Nickerson's ordering ($\mathscr{N}$) can hold is if the agent comes into the experiment thinking that *properties R and B are positively correlated*. This is a precise form of *confirmation bias*, which commits the agent to the (implicit) view that $\boxed{\sim B}$*-refutation is less probable than* $\boxed{B}$*-confirmation*. Of course, it is not a new idea that some sort of *confirmation bias* might be implicated in the Wason Task(s) [27]. But, this result establishes that the standard, raven-paradox-inspired Bayesian approaches to the Wason Task(s) are *committed* to the existence of confirmation bias in this precise, probabilistic-correlational sense.[27] What this reveals is that it is more difficult to *rationalize* the behavior/performance of actual subjects on the Wason Task(s) than one might have thought. While there might be some contexts in which this kind of confirmation bias is innocuous, we doubt it will be kosher *in general*. Moreover, we think our present results should also be of interest to those who are engaged (merely) in *explaining how it is possible* for a (coherent) Bayesian subject to exhibit the preferences over evidence-gathering strategies that actual subjects tend to exhibit in the Wason Task(s). From a "how possibly explanation" point of view, our result (Theorem 4) establishes that [taking (3′), (‡), *and* ($\star$) as background assumptions] it is only *possible* for a Bayesian *with a confirmation bias* to exhibit the kinds of preferences over evidence-gathering strategies that Nickerson's Bayesian analysis (and other Bayesian analyses [30]) require.

[26]Popper was infamous for claiming that *refutation* is *all important* for confirmation theory. Our assumption ($\star$) is *much weaker* (and more plausible) than Popperian falsificationism. All ($\star$) commits us to is the claim that refuting instances are *more powerful* than positive instances. Unlike Popper, we think positive instances *can be* highly probative, just not quite *as* probative as refuting instances.

[27]Oaksford & Chater [29] have a rather different probabilistic approach to the Wason Task(s). But, we suspect a similar "confirmation bias" result will be applicable to their approach as well (given the kinds of background assumptions we're making here). We don't have the space here, unfortunately, to delve into alternative Bayesian approaches to the Wason Task(s) that have appeared in the recent literature. But, see [30] for a nice discussion of the similarities between various recent probabilistic approaches to the Wason Task(s), including Nickerson's [26] and Oaksford & Chater's [29].

We have some other complaints about Nickerson's approach (*e.g.*, his choice of the difference measure d as his underlying measure of degree of confirmation), but we'll let those pass (presently).[28] In the next (and final) section, we'll discuss what Carnap might have said about the Wason Task, in light of Maher's recent work, and Nickerson's technique. Intriguingly, this will bring us back to Wason himself.

2.7. Carnap (*via* Maher) + Nickerson $\Longrightarrow$ Wason.

We'll close with a discussion of the following question: What happens when you take Maher's Carnapian probability models and apply them to the Wason Task, using Nickerson's technique? Answer: you get a *Wasonian* "normative" ordering!

Allow us to explain. Let $\Pr(\cdot \mid \top)$ be some conditional probability function drawn from Maher's [24] Carnapian models (and conditionalized only on *tautological* evidence). What happens when we calculate the (Nickersonian) expected confirmational powers $\mathcal{P}_\top$ of single-card strategies in the Wason Task, using these Carnapian probabilities? Answer: we are *forced* into the following *Wasonian* ordering:

($\mathscr{W}$) $\mathcal{P}_\top\left(\boxed{R}\right) \geq \mathcal{P}_\top\left(\boxed{\sim B}\right) > \mathcal{P}_\top\left(\boxed{B}\right)$.

That is, ($\mathscr{W}$) must be satisfied by *every* (tautological) Carnapian probability function $\Pr(\cdot \mid \top)$. This means Nickerson's ordering ($\mathscr{N}$) is incompatible with *any* (tautological/Hempelian) Carnapian approach to The Paradox of Confirmation.[29]

This brings us full circle, back to our Hempelian roots. Suppose that Maher's Carnapian models succeed at furnishing a (charitable) probabilistic reconstruction of a Hempelian "tautological" confirmation relation.[30] And, suppose that Nickerson is correct in his (empirical) conjecture that actual subjects rank $\boxed{B}$ *above* $\boxed{\sim B}$ in terms of expected confirmational power. Then, from the point of view of the traditional, Hempelian perspective on The Paradox of Confirmation, *Wason was right* in his negative assessment of his subjects' responses to his Task(s). That is, if these two suppositions are correct, then the empirical ordering ($\mathbb{O}$) violates the norms implied by the original ("logical") confirmation theories of Hempel and Carnap. We think this is an interesting (possible) way to vindicate Wason's normative intuitions, even in light of Humberstone's [20] interpretation-(II) of Wason's Task.

[28] In our *technical addendum*, we offer an alternative probabilistic analysis of our Wason Task(s).

[29] We suspect that *any* plausible "inductive-logical" theory of confirmation will view the ranking of $\boxed{B}$ over $\boxed{\sim B}$ as *mistaken*. Essentially, this is because all such theories will be committed to our assumption ($\star$), above. Indeed, ($\star$) is implied by all "inductive-logical" approaches to $\bar{c}$ we know of.

[30] As a matter of historical fact, this is exactly what Carnap [1, p. 472] was *trying* to do, with his notion of "initial confirmation". So, this is not a *historically* unreasonable supposition. Moreover, Carnap [1, p. xvi] does not seem to disapprove of Nickerson's choice of measure of confirmation (d).

References

[1] R. Carnap, *Logical Foundations of Probability*, Chicago University Press, Chicago, 1950 (second edition 1962).

[2] R.L. Dominowski, *Content effects in Wason's selection task*, in *Perspectives on thinking and reasoning: Essays in honour of Peter Wason*, S. Newstead and J. Evans (*eds.*), Psychology Press, 1995.

[3] J. Earman, *Bayes or bust?*, MIT Press, Cambridge, MA, 1992.

[4] B. Fitelson, *Studies in Bayesian confirmation theory*, Ph.D. thesis, University of Wisconsin–Madison (Philosophy), 2001, `http://fitelson.org/thesis.pdf`.

[5] ———, *Inductive Logic*, in *The Philosophy of Science: An Encyclopedia*, J. Pfeifer and S. Sarkar (*eds.*), Oxford: Routledge, 2005.

[6] ———, *The paradox of confirmation*, *Philosophy Compass* (B. Weatherson and C. Callender, *eds.*), Blackwell (online publication), Oxford, 2005, `http://fitelson.org/ravens.htm`.

[7] ———, *Goodman's 'New Riddle'*, *Journal of Philosophical Logic* **37** (2008), 613–643.

[8] B. Fitelson and J. Hawthorne, *How Bayesian confirmation theory handles the paradox of the ravens*, *Probability in Science* (Eells and Fetzer, *eds.*), *Boston Studies in the Philosophy of Science*, Vol. 284, pages 247–276, Springer, Dordrecht, 2010.

[9] I.J. Good, *The white shoe is a red herring*, *British J. for the Phil. of Sci.* **17** (1967), 322.

[10] ———, *The white shoe* qua *red herring is pink*, *British J. for the Phil. of Sci.* **19** (1968), 156–157.

[11] ———, *Good Thinking: The Foundations of Probability and its Applications*, University of Minnesota Press, Minneapolis, 1983.

[12] N. Goodman, *Fact, Fiction, and Forecast*, Harvard University Press, Cambridge, Mass., 1955.

[13] G. Harman, *Change in View: Principles of Reasoning*, MIT Press, 1988.

[14] J. Hawthorne, *Degree-of-Belief and Degree-of-Support: Why Bayesians Need Both Notions*, *Mind* **114** (2005) 277–320.

[15] ———, *Inductive Logic*, in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (*ed.*), 2010, `http://plato.stanford.edu/entries/logic-inductive/`.

[16] C. Hempel, *A purely syntactical definition of confirmation*, *JSL* **8** (1943), 122–143.

[17] ———, *Studies in the logic of confirmation*, *Mind* **54** (1945), 1–26, 97–121.

[18] ———, *The white shoe: no red herring*, *British J. for the Phil. of Sci.* **18** (1967), 239–240.

[19] C. Hempel & P. Oppenheim, *A definition of "degree of confirmation"*, *Phil. Sci.*, **12** (1945), 98–115.

[20] L. Humberstone, *Hempel Meets Wason, Erkenntnis* **41** (1994), 391–402.

[21] J. Keynes, *A Treatise on Probability*, Macmillan, London, 1921.

[22] J. MacFarlane, *In what sense (if any) is logic normative for thought?*, manuscript, 2004.

[23] P. Maher, *Inductive Logic and the Paradox of the Ravens*, *Philosophy of Science* **66** (1999), 50–70.

[24] ———, *Probability captures the logic of scientific confirmation*, in *Contemporary Debates in the Philosophy of Science* (C. Hitchcock, *ed.*), Blackwell, Oxford, 2004.

[25] C. McKenzie and L. Mikkelsen, *The Psychological Side of Hempel's Paradox of Confirmation*, *Psychonomic Bulletin & Review* **7** (2000), 360–66.

[26] Nickerson, R. *Hempel's Paradox and Wason's Selection Task: Logical and Psychological Puzzles of Confirmation*, *Thinking and Reasoning* **2** (1996), 1–31.

[27] ———, *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*, *Review of General Psychology* **2** (1998), 175–220.

[28] J. Nicod, *The logical problem of induction*, (1923) in *Geometry and Induction*, University of California Press, Berkeley, California, 1970.

[29] Oaksford, M. and Chater N. *A Rational Analysis of the Selection Task as Optimal Data Selection*, *Psychological Review* **101** (1994), 608–631.

[30] ———, *Optimal data selection: Revision, review, and reevaluation*, *Psychonomic Bulletin & Review* **10** (2003), 289-318.

[31] W.V.O. Quine, *Natural kinds*, in *Ontological Relativity and Other Essays*, Columbia U. Press, 1969.

[32] F.P. Ramsey, *Truth and Probability*, in *The foundations of mathematics and other logical essays*, Kegan Paul, London, 1931.

[33] Royall, R. *Statistical Evidence: A Likelihood Paradigm*, Chapman & Hall/CRC, 1997.

[34] I. Scheffler, *Anatomy of Inquiry*, Knopf, New York, 1963.

[35] I. Scheffler and N. Goodman, *Selective Confirmation and the Ravens: A Reply to Foster*, *The Journal of Philosophy*, **69** (1972), 78-83.

[36] R. Sylvan and R. Nola, *Confirmation without paradoxes*, in *Advances in Scientific Philosophy* (G. Schurz and G. Dorn, *eds.*), Rodopi, Amsterdam/Atlanta, 1991.

[37] P. Vranas, *Hempel's raven paradox: a lacuna in the standard Bayesian solution*, *British Journal for the Philosophy of Science*, Vranas **55** (2004): 545–560.

[38] Wason, P. and D. Shapiro. *Natural and Contrived Experience in a Reasoning Problem*, *Quarterly Journal of Experimental Psychology* **23** (1971), 63–71.

[39] T. Williamson. *Knowledge and its Limits*. Oxford University Press, Oxford, 2002.

PHILOSOPHY DEPT., RUTGERS UNIVERSITY, 1 SEMINARY PLACE, NEW BRUNSWICK, NJ 08901–1107.
E-mail address: `branden@fitelson.org`
URL: `http://fitelson.org/`

PHILOSOPHY DEPT., UNIVERSITY OF OKLAHOMA, 455 W. LINDSEY, NORMAN, OK 73019-2006.
E-mail address: `hawthorne@ou.edu`
URL: `http://faculty-staff.ou.edu/H/James.A.Hawthorne-1/`