Probability, Confirmation, and the Conjunction Fallacy

Vincenzo Crupi[*] Branden Fitelson[†] Katya Tentori[‡]

July, 2007

[*] Department of Arts and Design (University IUAV of Trento)
and Cognitive Psychology Laboratory, CNRS (University of Aix-Marseille I)

[†] Department of Philosophy (University of California, Berkeley)

[‡] Department of Cognitive Sciences and Education and CIMeC
(University of Trento)

Probability, Confirmation, and the Conjunction Fallacy[1]

Abstract. The conjunction fallacy has been a key topic in debates on the rationality of human reasoning and its limitations. Despite extensive inquiry, however, the attempt of providing a satisfactory account of the phenomenon has proven challenging. Here, we elaborate the suggestion (first discussed by Sides *et al*., 2001) that in standard conjunction problems the fallacious probability judgments experimentally observed are typically guided by sound assessments of *confirmation* relations, meant in terms of contemporary Bayesian confirmation theory. Our main formal result is a confirmation-theoretic account of the conjunction fallacy which is proven *robust* (i.e., not depending on various alternative ways of measuring degrees of confirmation). The proposed analysis is shown distinct from contentions that the conjunction effect is in fact not a fallacy and is compared with major competing explanations of the phenomenon, including earlier references to a confirmation-theoretic account.

1. Introduction: probability and confirmation in inductive logic

Inductive logic may be seen as the study of how a piece of evidence *e* affects the credibility of a hypothesis *h*. Within contemporary epistemology, a major perspective on this issue is provided by Bayesianism. Early Bayesian theorists, such as Carnap (1950), proposed the conditional (or posterior) probability of *h* on *e* as an appropriate formalisation of the basic inductive-logical relationship between evidence and hypothesis. This account, however, led to counterintuitive consequences and conceptual contradictions, emphasized in a now classical debate (see Popper, 1954). Later on, Carnap himself came to a fundamental distinction between the notions of *firmness* and *increase in firmness* of a hypothesis *h* in the light of evidence *e*, and reached the conclusion that the posterior of *h* could be taken as accounting for the former concept, but not the latter (Carnap, 1962). In fact, the credibility of a hypothesis *h* (e.g., a diagnosis) may *increase* as an effect of evidence *e* (e.g., a positive result in a diagnostic test) and still remain relatively *low* (for instance, because the concerned disease is very rare); similarly, *e* might *reduce* the credibility of *h* while leaving it rather *high.* As simple as it is, this distinction is of the utmost importance for contemporary Bayesianism.

Epistemologists and inductive logicians working within the Bayesian framework have proposed a plurality of models to formalise and quantify the notion of *confirmation*, meant in terms of Carnap's *increase in firmness* brought by *e* to *h* (or, equivalently, as the *inductive strength* of the argument from *e* to *h*). Each proposal maps a pair of

statements e,h on a real number which is positive in case $p(h|e) > p(h)$ (i.e., when e *confirms* h), equals 0 in case $p(h|e) = p(h)$ (i.e., when e is *neutral* for h), and is negative otherwise (i.e., when e *disconfirms* h). *Table 1* reports a representative sample of alternative Bayesian measures of confirmation discussed in the literature (see Festa, 1999; Fitelson, 1999).

Table 1. Alternative Bayesian measures of confirmation.

$D(h,e) = p(h \mid e) - p(h)$	(Carnap, 1950; Eells, 1982)
$R(h,e) = \ln\left[\frac{p(h \mid e)}{p(h)}\right]$	(Keynes, 1921; Milne, 1996)
$L(h,e) = \ln\left[\frac{p(e \mid h)}{p(e \mid \neg h)}\right]$	(Good, 1950; Fitelson, 2001)
$C(h,e) = p(h \,\&\, e) - p(h) \times p(e)$	(Carnap, 1950)
$S(h,e) = p(h \mid e) - p(h \mid \neg e)$	(Christensen, 1999; Joyce, 1999)
$Z(h,e) = \begin{cases} \frac{p(h \mid e) - p(h)}{1 - p(h)} & \text{if } p(h \mid e) \geq p(h) \\ \frac{p(h \mid e) - p(h)}{p(h)} & \text{otherwise} \end{cases}$	(Crupi, Tentori & Gonzalez, 2007)

It is well known that $p(h|e)$ and $c(h,e)$ – where c stands for any of the Bayesian measures of confirmation listed above – exhibit remarkably different properties. One such difference will play a crucial role in what follows. It amounts to the following fact:

(1) $h_1 \models h_2$ implies $p(h_1|e) \leq p(h_2|e)$ but does *not* imply $c(h_1,e) \leq c(h_2,e)$

To illustrate, consider the random extraction of a card from a standard deck, and let e, h_1 and h_2 be statements concerning the drawn card, as follows:

e = “black card”
h_1 = “picture of spades”
h_2 = “picture card”

Notice that, clearly, $h_1 \models h_2$, so the probability of the former cannot exceed that of the latter, even conditionally on e. In fact, by the standard probability calculus, $p(h_1|e) = 3/26 < 6/26 = p(h_2|e)$. However, the reader will concur that e positively affects the credibility of h_1 while leaving that of h_2 entirely unchanged, so that $c(h_1,e) > c(h_2,e)$. This is because $p(h_1|e) = 3/26 > 3/52 = p(h_1)$, whereas $p(h_2|e) = 6/26 = 12/52 = p(h_2)$. Examples such as this one effectively highlight the crucial conceptual distinction between probability and confirmation.

2. Probability and confirmation in the psychology of induction

The consideration of normative models of reasoning is often relevant when interpreting empirical studies of human cognition. Unfortunately, with few notable exceptions (e.g., Sides, 2001), a clear distinction between probability and confirmation is seldom spelled out in psychological analyses of human inductive reasoning, so that the properties of the former are sometimes unduly attributed to the latter.

To illustrate the point for the present purposes, we will rely on a touchstone study in the psychology of inductive reasoning, carried out by Osherson *et al*. (1990), whose participants were presented with arguments composed by statements attributing "blank" predicates to familiar biological categories (such as "mice"). Blank predicates are meant as "indefinite in their application to given categories, but clear enough to communicate the kind of property in question" (Lo *et al*., 2002). For example, "use serotonin as a neurotransmitter" is a blank predicate, for most reasoners are unaware of the animals to which it does or does not apply yet it clearly refers to a biological property. In one of Osherson's *et al*. (1990) experiments, subjects faced a pair of arguments of the following form (where the statements above and below the bar serve as premise and conclusion, respectively):

(e) robins have property P	(e) robins have property P
---	---
(h_1) all birds have property P	(h_2) ostriches have property P

When asked to "choose the argument whose facts provide a better reason for believing its conclusion", a robust majority (65%) chose argument e,h_1. Notice that these instructions may be legitimately interpreted as eliciting an (ordinal) judgment of confirmation, i.e., in our terms, a ranking of $c(h_1,e)$ and $c(h_2,e)$. Argument e,h_1, however, also scored a higher rating by most subjects from a different group when asked to "estimate the probability of each conclusion on the assumption that the respective premises were true", i.e., $p(h_1|e)$ and $p(h_2|e)$. Osherson *et al*. (1990) convincingly argue that these results are connected to the fact that robins are perceived as highly typical birds while ostriches are not.

The former results, taken as a whole, are commonly labelled a "fallacy" in the psychological literature on inductive reasoning, on the basis that $h_1 \models h_2$ (see, for instance: Gentner & Medina, 1998, p. 283; Heit, 2000, p. 574; Sloman & Lagnado, 2005, p. 105). Yet, the undisputed mathematical fact expressed by statement (1) above implies a more articulated diagnosis: a fallacy is certainly there when the posteriors of h_1 and h_2, respectively, are at issue; it is not necessarily so, however, if the two arguments are assessed by their inductive strength, i.e., in terms of confirmation. Moreover, coherent probability assignments exist by which *all* the confirmation

measures listed above do imply $c(h_1,e) > c(h_2,e)$. To see this, it suffices to apply a method of analysis of categorical arguments proposed by Heit (1998) and consider the probability assignments reported in *Table 2*.

Table 2. Possible probability assignments concerning e ("robins have property P"), h_1 ("all birds have property P") and h_2 ("ostriches have property P").

conjunction n.		$p(\text{conj}_i)$	$p(e\|\text{conj}_i)$	$p(\text{conj}_i\|e)$
conj_1:	e & h_1	.20	1	.57
conj_2:	e & $\neg h_1$	.15	1	.43
conj_3:	$\neg e$ & h_1	0	0	0
conj_4:	$\neg e$ & $\neg h_1$	.65	0	0
conj_5:	e & h_2	.22	1	.63
conj_6:	e & $\neg h_2$	.13	1	.37
conj_7:	$\neg e$ & h_2	.13	0	0
conj_8:	$\neg e$ & $\neg h_2$	.52	0	0

The table does not contain any inconsistency and has been built to convey the following statements:

- $p(e) = p(h_2)$, for, insofar as P is a blank predicate, it seems reasonable to treat it as if it expressed a randomly selected biological property, which is equally likely to pertain to robins as to any other kind of birds, such as ostriches;
- $p(h_1) < p(h_2)$, since the former implies the latter (not the converse);
- $p(e\&h_1) < p(e\&h_2)$, since the former implies the latter (not the converse), but the difference between the two is minor, for robins are highly typical birds but ostriches are not, therefore the properties shared by robins and ostriches are virtually only those shared by robins and birds.

By the values in *Table 2*, it can be computed that $p(h_1) = .2$, $p(h_2) = .35$, $p(h_1|e) = .57$ and $p(h_2|e) = .63$. On these conditions, it is easy to show that, for any of the confirmation measures in *Table 1*, $c(h_1,e) > c(h_2,e)$, which reflects precisely the ranking exhibited by participants' responses. (Computational details omitted.) Importantly, this result does not depend on a selective choice of the value of priors such as $p(h_1)$, since a similar table may be constructed wherein, for instance, $p(h_1) = .5$. Thus, a Bayesian account of confirmation may in fact not only *be consistent* with the observed ranking of inductive strength (an immediate consequence of statement (1) above), but even *imply* it under plausible assumptions. And it can do that *robustly* (in the sense of Fitelson, 1999), i.e., independently of the choice of a particular confirmation measure among those listed in *Table 1*.

By the foregoing analysis, one reading of the participants' responses naturally arises: possibly, even when judging posterior probabilities, people's evaluations were guided by assessments of the degree of confirmation provided by e to h_1 and h_2, respectively.

Notice that this *does not imply in any way* that participants in the probability task *meant* to judge something else other than probability, thus consciously giving the experimental stimuli an unanticipated interpretation. Consequently, the hypothesis that the appreciation of confirmation relations might have driven those probability judgments does *not* amount to a pragmatically-inspired contention of the diagnosis that an error did occur in the probability task itself. Rather, it is an explicative conjecture as to why it may have in fact occurred. (We'll come back to this point in detail later on. See paragraph 4.)

Another important study by Lagnado & Shanks (2002) offers further evidence in support of this conjecture. The authors referred to confirmation measure *S* in *Table 1* as a measure of the "predictiveness" of *e* with regards to *h*. Then, they manipulated this quantity in a sophisticated learning task involving various symptoms (e.g., *e* = "high muscle tension") and a hierarchically structured set of possible diagnostic hypotheses, some of them (e.g., h_1 = "flu") being implied by others (e.g., h_2 = "Chinese flu"). According to their results, the fact that in the learning task $S(h_1,e) > S(h_2,e)$ accounted for the occurrence of the counternormative pattern of judgments $p(h_1|e) > p(h_2|e)$ in a subsequent probability rating task. Apparently, these participants' incoherent probability judgments rested in fact on (sound) assessments of confirmation (predictiveness).

In what follows, the working hypothesis that in certain circumstances reported assessments of probability may reflect the appreciation of confirmation relations will be applied in detail to one of the most widely known and discussed phenomenon in the study of human reasoning: the "conjunction fallacy".

3. Linda, the patient and Bjorn Borg: a unified confirmation-theoretical analysis

A number of studies have reported that, in the presence of some available evidence (*e*), people may judge a conjunction of hypotheses (h_1&h_2) as more probable than one of its conjuncts, contrary to the principle of probability known as the "conjunction rule". Three examples taken by the seminal work of Tversky & Kahneman (1983) will serve as illustration for our purposes.

- When faced with the description of a character, Linda, 31 years old, single, outspoken and very bright, with a major in philosophy, concerns about discrimination and social justice and an involvement in anti-nuclear demonstrations (*e*), most people ranked "Linda is a bank teller and is active in the feminist movement" (h_1&h_2) as more probable than "Linda is a bank teller" (h_1).
- Given the description of the clinical case of a 55-old woman with a pulmonary embolism documented angiographically 10 days after a cholecystectomy (*e*), a large

majority of physicians judged that the patient would more likely experience emiparesis and dyspnea (h_1&h_2) than emiparesis (h_1).

- Asked soon after Borg's victory of his fifth consecutive Wimbledon in 1980 (e) (when, as Tversky & Kahneman remarked, "Borg seemed extremely strong", p. 31), the majority of participants asserted that, having reached the final in the 1981 edition, Borg would have more probably lost the first set but won the match (h_1&h_2) than lost the first set (h_1).

The above examples represent a whole class of findings about conjunction problems sharing a distinctive set of common traits:

(i) e is negatively (if at all) correlated with h_1;
(ii) e is positively correlated with h_2, *even conditionally on* h_1;
(iii) h_1 and h_2 are mildly (if at all) negatively correlated.

It is presently submitted that a unified account of probabilistic fallacious judgments in classical conjunction problems could be found on the basis of the notion of confirmation: participants may in fact have a tendency to rely on assessments of confirmation when judging probabilities. More precisely, the hypothesis is that, on conditions (i)-(iii), most participants may depart from the relevant probabilistic relationship between $p(h_1\&h_2|e)$ and $p(h_1|e)$ because of the perception that $c(h_1\&h_2,e) > c(h_1,e)$. (See paragraph 5. for a detailed discussion of earlier suggestions in this vein.)

The following theorem (proven in the Appendix) shows that, for *any* choice among major alternative confirmation measures, appropriate confirmation-theoretic renditions of (i) and (ii) above are sufficient to imply the ordering $c(h_1\&h_2,e) > c(h_1,e)$.

Theorem. For any Bayesian measure of confirmation c among D, R, L, C, S and Z,
if (i*) $c(h_1,e) \leq 0$ and (ii*) $c(h_2,e|h_1) > 0$,
then $c(h_1\&h_2,e) > c(h_1,e)$.[2]

A plurality of plausible cognitive processes may converge on the judgment that $c(h_1\&h_2,e) > c(h_1,e)$. First of all, notice that the appreciation of e fostering the credibility of h_2 but not h_1 (i.e., e confirming the former but not the latter) seems quite straightforward in standard conjunction problems such as Linda, the patient and Borg. Given that, people's judgment about the effect of e on $h_1\&h_2$ may reflect the estimation of an average (either weighted or simple) of the (positive) perceived strength of argument e,h_2 and the (negative or null) perceived strength of e,h_1.[3] Also, variants of an "anchoring and adjustment" process (Tversky & Kahneman, 1974), by which the perceived strength of one of the arguments is subsequently adjusted towards the other, would produce the same outcome. The point of the present analysis is that the result of such a line of thought, incoherent as a probability ranking (and thus a genuine error given the experimental task), could be accounted for on a confirmation-theoretic reading. In fact, this account fleshes out and extends the otherwise esoteric remark by

Tversky and Kahneman themselves that "*feminist bank teller* is a better hypothesis about Linda than *bank teller*" (1983, p. 45). It is, we submit, in the sense that it is *better confirmed* by Linda's description. The same occurs with the other examples discussed. In such conditions, "the answer to a question [probability] can be biased by the availability of an answer to a cognate question [confirmation]" (Tversky & Kahneman, 1983, p. 47, square brackets added).

4. A "fallacious fallacy"?

The conjunction fallacy has become a key topic in debates on the rationality of human reasoning and its limitations (see Stich, 1990, Kahneman & Tversky, 1996, and Gigerenzer, 1996, among others). One reaction has been the claim that the experimental evidence on conjunction problems has not demonstrated the occurrence of a reasoning fallacy after all. It is then crucial to discuss such a claim and keep it distinct from the implications of the present analysis.

Recurrent worries have been inspired by the pragmatics of communication in experimental settings. According to this line of argument, experimental subjects might have in fact interpreted the isolated conjunct "h_1" as "h_1 *and not-*h_2" (see Morier & Borgida, 1984; MacDonald & Gilhooly, 1986; Politzer & Novack, 1991; Dulany & Hilton, 1991; Hilton, 1995), or they might have read the ordinary-language conjunction "and" as a disjunction (see Mellers, Hertwig & Kahneman, 2001). The results of recent experiments devised to investigate these possible sources of confound suggest that the first one of them might have contributed to the size of the effect in earlier documentations of the phenomenon (Sides *et al*., 2001; Bonini, Tentori & Osherson, 2004; Tentori, Bonini & Osherson, 2004). However, these studies have also clearly shown that the conjunction fallacy phenomenon persists despite such conversational implicatures being strongly discouraged or otherwise controlled for.

It has then been observed and documented that lay people may rephrase the term "probability" in disparate ways (Hertwig & Gigerenzer, 1999). It is also well known, however, that participants, when debriefed, do *not* usually defend their response that the conjunction is more probable on the basis of an alternative meaning of "probable". Rather, they normally concede making an error (and seem to experience some spontaneous regret for it). Of course, this is taken by many as the hallmark of cognitive illusions, and we are not aware of any compelling argument in defense of the rationality of the conjunction effect which handles this circumstance conveniently. Furthemore, the choice of a conjunction over a single conjunct have been documented under *betting* instructions, wherein the mathematical probability of winning is the uncontroversial criterion for rational behaviour and yet the term "probability" itself is not even mentioned (Tversky & Kahneman, 1983; Sides *et al*., 2001; Bonini, Tentori & Osherson, 2004).

It has also been claimed that frequency formats make cognitive illusions (among which the conjunction fallacy itself) "largely disappear" (Gigerenzer, 1996, p. 595). Yet there is evidence that the conjunction fallacy persists under a frequentist presentation and that such a presentation does not even always affect its prevalence as compared to a standard probability format (Sloman *et al*., 2003; Tentori, Bonini & Osherson, 2004). Even more to the point, for our purposes, is the remark that the frequentist approach, whatever its merits, "does not explain why people make the errors in the first place under a probability format" (Lagnado & Shanks, 2002, p. 108).[4] Likewise, this does not seem to be explained by other findings concerning conditions which may increase conformity to the conjunction rule (e.g., ratings *vs*. ranking tasks; see Hertwig & Chase, 1998, and Sloman *et al*., 2003).

A more theoretically-oriented defense of the rationality of human judgment in standard conjunction experimental problems has been recently advocated by Bovens & Hartmann (2003, pp. 85-88) and Hintikka (2004). Briefly put, the proposal is the following. Suppose "Linda is a bank teller" (h_1) and "Linda is a feminist bank teller" (h_1&h_2) are reports of two distinct sources of information s_1 and s_2 which are not perfectly reliable. Linda's description *e* may well suggest that source s_1 is *less reliable* than s_2. But then, probability theory *is* consistent with the statement that the probability of h_1 *conditional on the relatively low reliability* of s_1 is lower than the probability of h_1&h_2 *conditional on the relatively high reliability* of s_2. It is submitted that *this* is what participants' responses express.

It has been observed, however, that standard experimental stimuli are completely silent about h_1 and h_1&h_2 being reports of two distinct sources of information (see Levi, 2004, p. 37; Olsson, 2005, p. 292). We would add that the plausibility of the above reconstruction is shown even more problematic by the conjunction fallacy occurring in problems (such as either the patient or Borg) involving hypotheses about *future* events. For one has to make the additional assumption that in such cases participants interpret the task as concerning *forecasts* h_1 and h_1&h_2 as made by two distinct predictors, which again are never mentioned in the experimental scenarios.

More generally, in denying the fallacious character of the conjunction effect, the proponents of the latter account seem to have shared with other authors (e.g., Levi, 2004) the endorsement of the following line of argument: a formally defined attribute is identified which, in certain conditions, would appropriately rank h_1&h_2 over h_1; *thereby* the conclusion is *immediately drawn* that the conjunction effect is in fact a "fallacious fallacy" (Hintikka, 2004), viz., that "there need not be anything fallacious or otherwise irrational about the conjunction effect" (p. 30). However, pending an independent argument to the effect that in standard probabilistic (and betting) conjunction tasks participants are rationally justified in evaluating something else other than probabilities $p(h_1|e)$ and $p(h_1\&h_2|e)$, we see this inference as spurious.

5. Explaining the fallacy

As argued above, in our view the diagnosis of the conjunction effect reflecting a reasoning fallacy stands. What is at issue are its determinants in human cognition. Accordingly, in what follows we will compare our account with some important alternative explanations. It will also be pointed out in which respects the present analysis improves on earlier references to a confirmation-theoretic account.

A reading of the conjuntion fallacy effect has been proposed within *support theory* (Tversky & Koehler, 1994; Brenner, Koehler, & Rottenstreich, 2002). Support theory is a formal framework departing from classical probability theory and devised as a descriptive account of subjective probability assessments. It models subjective probability as depending on a newly introduced psychological construct which is labelled the *support* associated with a given hypothesis and is informally interpreted as "the strength of evidence in favor of this hypothesis" (Tversky & Koehler, 1994, p. 445). From the postulated properties of the support function, a critical (non-normative) tenet of the theory is derived (also labelled *unpacking principle*), i.e., the *subaddivity* of the judged probability of a hypothesis *h* with regards to the judged probabilities of a set of mutually exclusive hypotheses whose disjunction is logically equivalent to *h*. The relevant instantiation of this statement would amount to the following disequality:

(2) $p(h_1|e) \leq p(h_1\&h_2|e) + p(h_1\&\neg h_2|e)$

Expression (2) says, for instance, that given Linda's character the judged probability of her being a bank teller may be lower than the judged probability of her being a feminist bank teller plus the judged probability of her being a non-feminist bank teller. (2) is inconsistent with the conjunction rule and compatible with its violation. However, the conjunction fallacy reflects a significantly more extreme pattern than simple subadditivity, i.e.:

(3) $p(h_1|e) < p(h_1\&h_2|e)$

To the best of our knowledge, although consistent with pattern (3), support theory does not provide grounds to predict its occurrence under independently specified conditions. Similar difficulties arise with other algebraic models which, although consistent with the conjunction fallacy effect, can account for the phenomenon only by letting quite a few free parameters be determined from the data to be explained (see, for instance, Birnbaum, Anderson, & Hynan, 1990; and Massaro, 1994). The confirmation-theoretic account we present, by contrast, does specify a set of conditions on which the conjunction fallacy effect is expected, which may be subject to independent empirical control by the elicitation of judgments involving the confirmation relations among *e*, h_1 and h_2 (for more on this, see the next section).

A more empirically grounded approach has been taken by Shafir, Smith & Osherson (1990), elaborating on Tversky and Kahneman's original hypothesis of the "representativeness heuristic". The authors of this study have collected "typicality ratings" of Linda's character relative to the single category "bank teller" and the conjoint category "feminist bank teller" and interpreted such ratings as reflecting intuitive assessments of the probability of *e* given h_1 and $h_1\&h_2$, respectively. In Linda's problem, and in a set of similar cases, such typicality ratings have proven reliable predictors of the conjunction fallacy effect. However, the "inverse probability" account, i.e., the explanatory hypothesis of people's assessment of posteriors $p(h_1|e)$ and $p(h_1\&h_2|e)$ by an evaluation of the likelihoods $p(e|h_1)$ and $p(e|h_1\&h_2)$ is not easily extended to the medical or the Borg cases above. In fact, this would imply the rather cumbersome judgmental strategy of focussing on the probability of the known clinical frame and Borg's past record, respectively, *conditional on future (hypothetical) events* such as the manifestation of certain symptoms or the outcome of a match. The problem of future events does not arise in a confirmation-theoretic account, however, insofar as confirmatory or disconfirmatory impact of a piece of evidence *e* can be, and often is, naturally assessed whatever the state of affairs (either past, present or future) to which the concerned hypothesis *h* refers.

The most prominent antecedent of the present treatment is the neat analysis carried out by Sides *et al*. (2001) at the beginning of their paper. In our view, however, although important, such an analysis has the limitation of being measure-dependent, i.e., not *robust*. In fact, it only refers to the "ratio measure" (measure *R* in *Table 1*). The crucial point here is that alternative Bayesian confirmation measures do *not* generally agree in their implied rankings; quite on the contrary, they are known to disagree in many crucial cases. (For instance, some measures only – precisely *R* and *C* in our list – have the disputable consequence that *e* will confirm *h* exactly to the same extent as *h* confirms *e* for any possible choice of *e* and *h*. See Carnap, 1962, § 67, and Eells & Fitelson, 2002.) This being so, the extrapolation from a non-robust to a robust (i.e., *not* measure-dependent) account is generally unwarranted, and the existence of the latter is far from a trivial issue in many cases (Fitelson, 1999, provided abundant evidence for the relevance of measure-dependence in a wide range of epistemological matters). Furthermore, the adequacy of the very measure *R* on which Sides's *et al*. (2001) analysis is based has been found questionable on both normative and empirical grounds (see Eells & Fitelson, 2002; Tentori *et al*., 2007; and Crupi, Tentori & Gonzalez, 2007). The theorem on which the present account is centred removes the foregoing worries by showing that, for any choice among major alternative confirmation measures, a few conditions which apparently hold for commonly used conjunction problems are sufficient to imply the ordering $c(h_1\&h_2,e) > c(h_1,e)$.

In the discussion section of their experimental study on "predictiveness" and probability judgment, Lagnado & Shanks (2002) also considered a confirmation

measure – *S* in our list – and remarked that *S*(feminist bank teller, Linda) may well be higher than *S*(bank teller, Linda) (p. 108). They did not, however, formally prove that this will in fact be the case under a defined set of general conditions. Similarly, Levi (2004) has claimed, but not formally proven, that the "difference" and "ratio" measures (*D* and *R*) "are to be expected to rank 'feminist bank teller' over 'bank teller'" (p. 38). Finally, Hertwig & Chase (1998), in their extensive empirical inquiry on the issue, also referred to a Bayesian confirmation measure due to Nozick (1981) – i.e., $p(e|h) - p(e|\neg h)$. It should be pointed out that our formal result can readily be extended to Nozick's measure (proof omitted). However, once again, relying on one particular confirmation measure is not a matter of taste and has consequences: normative and descriptive limitations of Nozick's measure have been pointed out by Eells & Fitelson (2002), Tentori *el al*. (2007) and Crupi, Tentori & Gonzalez (2007). Moreover, Hertwig & Chase (1998) only investigated Linda's case and explicitly endorsed the assumption that "participants' probability judgments are conditioned on the hypotheses rather than on the evidence" (p. 329), i.e., the inverse-probability hypothesis. Both points have been discussed above.

6. Concluding remarks

The present contribution is best seen in the framework of a view of cognitive biases arising from an overarching mechanism of "attribute-substitution" (Kahneman & Frederick, 2002). In the foregoing discussion, we tried to motivate our conjecture that (Bayesian) confirmation may be a better candidate surrogate attribute for probability as compared to competitors such as support and representativeness/typicality meant as inverse probability. This is, we submit, because alternative explanations are either less well specified or working properly only in a subset (typically, Linda's case) of the range of findings accounted for by our proposal, or both.[5] Of course, it is of interest that the probability-biasing surrogate attribute which is proposed *is* indeed rationally relevant in other contexts, as shown by extensive work in epistemology and related fields (see, for instance, Good, 1950, 1983, Joyce, 2004, and Fitelson, 2007). Yet we advance no contention here about the reality of the conjunction fallacy itself. In this perspective, the conjunction fallacy may be seen as a case of content prevailing over form. The suggestion is that, in standard conjunction experimental problems, content favours the assessment of confirmation-theoretic relationships among *e*, h_1 and h_2 to the detriment of the appreciation that, whatever h_1 and h_2 may be, any possible state of affairs satisfying (i.e., making true) $h_1 \& h_2$ also satisfies h_1, so that the probability of the former cannot possibly exceed that of the latter.

A limitation of the present account is, of course, that we are not presenting new empirical evidence in favour of it. We do think, however, that relevant experimental inquiries can be devised.

A first preliminary test could address our hypothesis that, in classical conjunction scenarios involving e, h_1 and h_2, people typically perceive confirmation-theoretic relationships as assumed, and most notably the impact of e on h_1 as negative (or null) and the impact of e on h_2 as positive. By an appropriate procedure eliciting judgments of confirmation (see, for instance, Tentori *et al.*, 2007; Tentori, Crupi, & Osherson, 2007), this could be checked for in the case of Linda as in the patient scenario, and presumably in a conveniently updated variant of the Borg case as well. Furthermore, our discussion immediately implies that, in this kind of conjunction problems, explicitly elicited assessments of $c(h_1\&h_2,e)$ and $c(h_1,e)$ should mirror the observed responses when evaluations of $p(h_1\&h_2|e)$ and $p(h_1|e)$ are requested.

More subtle tests would involve a quantitative refinement of our analysis. Insofar as the conjunction fallacy effect is supposed to be fostered by e having a negative (or null) impact on h_1 and a positive impact on h_2, it seems natural to expect the difference between mean ratings of $c(h_2,e)$ and of $c(h_1,e)$ to be positively correlated with the percentage of conjunction errors. This could be put to empirical test by means of a series of variants of a classical conjunction problem where $c(h_2,e)$ is manipulated, other things being kept constant. For instance, one may have several variants of the Linda scenario differing only for h_2 (e.g., say, no-global activist *vs*. poetry reader) and such that different (positive) mean ratings of $c(h_2,e)$ are obtained. Then, the higher the mean rating of $c(h_2,e)$, the higher the percentage of conjunction errors should be in the corresponding standard probabilistic task.

In a more theoretical vein, noticing that a perfectly Bayesian agent would never entertain inconsistent probabilities, one might find odd that the notion of Bayesian confirmation be invoked to account for a probabilistic fallacy. We do not think, however, that this concern is well-grounded. Indeed, we suspect that it rests on the misunderstanding of an alleged "supervenience" of the notion of confirmation on that of probability.

There is no question that, as a matter of historical fact, the standard formal treatment of probability reached an established form long ago, and thus served as a conceptual and technical basis for theories of confirmation. Formally, however, the relationship between the two notions is rather symmetric: simply, they mathematically constrain each other. Moreover, there is evidence that intuitive assessments of confirmation can be elicited directly, that – at least in some contexts – people can appropriately distinguish probability and confirmation and that their judgments satisfy, to a significant extent, the formal relationships between the two notions (see Tentori *et al.*, 2007; Crupi, Tentori, & Gonzalez, 2007). Indeed, an intriguing aspect of the results reported in Tentori *et al*. (2007) even suggests that the experimental problems discussed here may not be the only cases in which human reasoners judge confirmation more appropriately than probability. In this study it has been shown that, in an urn setting, normatively appealing Bayesian confirmation measures (such as measure L in our list) were better

predictors of elicited confirmation judgments when degrees of confirmation were computed by objective probabilities rather than subjectively judged ones, the latter having been found prone to well-known biases (in particular, “conservative” assessments of posteriors; see Edwards, 1968, and Slovic & Lichtenstein, 1971).

It is legitimate to ask whether general necessary and sufficient conditions can be specified under which the assessment of confirmation as a surrogate attribute for probability is expected to occur in human judgment. We see this as a fascinating issue for which further research is needed. Historically and theoretically, the relationships between the two concepts have proven rather subtle, and the same may turn out to be the case on the psychological level. In any event, the notion of confirmation has been an important conceptual tool in the normative analysis of inductive reasoning. In our opinion, the same could obtain in the descriptive study of such kind of reasoning (where it has not attracted comparable attention), and in the assessment of the relationships between the two.

References

Birnbaum, M.H., Anderson, C.J., & Hynan, L.G. (1990). Theories of bias in probability judgment". In J.P. Caverni, J.M. Fabre, & M. Gonzalez (eds.), *Cognitive biases* (pp. 477-499). North Holland: Elsevier.

Bovens, L. & Hartmann, S. (2003). *Bayesian epistemology*. Oxford, UK: Oxford University Press.

Brenner, L.A., Koehler, D.J., & Rottenstreich, Y. (2002). Remarks on support theory: Recent advances and future directions. In T. Gilovich, D. Griffin, & D. Kahneman (eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 489-509). New York: Cambridge University Press.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: University of Chicago Press.

Carnap, R. (1962). Preface. In *Logical foundations of probability* (2nd ed.). Chicago: University of Chicago Press.

Christensen, D. (1999). Measuring confirmation. *Journal of Philosophy*, *96*, 437– 461.

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues". *Philosophy of Science* (forthcoming).

Dulany, D.E. & Hilton, D.J. (1991). Conversational implicature, conscious representation and the conjunction fallacy. *Social Cognition*, *9*, 85-110.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (ed.), *Formal Representation of Human Judgment* (pp. 17-52). New York: Wiley.

Eells, E. & Fitelson, B. (2002). Symmetries and asymmetries in evidential support. *Philosophical Studies*, *107*, 129-142.

Eells, E. (1982). *Rational decision and causality*. Cambridge, UK: Cambridge University Press.

Fantino, E., Kulik, J., Stolarez-Fantino, S., & Wright, W. (1997). The conjunction fallacy: A test of averaging hypotheses. *Psychonomic Bulletin & Review*, *4*, 96-101.

Festa, R. (1999). Bayesian confirmation. In M. Galavotti & A. Pagnini (eds.), *Experience, reality, and scientific explanation* (pp. 55–87). Dordrecht: Kluwer.

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

Fitelson, B. (2001). A Bayesian account of independent evidence with applications. *Philosophy of Science*, *68*, S123-S140.

Fitelson, B. (2007). Likelihoodism, Bayesianism, and relational confirmation. *Synthese*, *156*, 473-489.

Gentner, D. & Medina, J. (1998). Similarity and the development of rules. *Cognition, 65,* 263-297.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky. *Psychological Review*, *103*, 592-596.

Good, I.J. (1950). *Probability and the weighing of evidence*. London: Griffin.

Good, I.J. (1983). *Good thinking*. Minneapolis: University of Minnesota Press.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (eds.), *Rational models of cognition* (pp. 248-274). New York: Oxford University Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7*, 569-592.

Hertwig, R. & Chase, V.M. (1998). Many reasons or just one: How response mode affects reasoning in the conjunction problem. *Thinking & Reasoning*, *4*, 319-352.

Hertwig, R. & Gigerenzer, G. (1999). The "conjunction fallacy" revised: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*, 275-305.

Hilton, D.J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin*, *118*, 248-271.

Hintikka, J. (2004). A fallacious fallacy? *Synthese*, *140*, 25-35.

Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge (UK): Cambridge University Press.

Joyce, J. (2004). Bayes's theorem. In E.N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (Summer 2004 Edition). URL = http://plato.stanford.edu/archives/sum2004/entries/ bayes-theorem/

Kahneman, D. & Frederick, S. (2002). Representativeness revised: Attribute substitution in intuitive judgment. In Gilovich, T., Griffin, D., & Kahneman, D. (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press, 49-81.

Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*, 582-591.

Keynes, J. (1921). *A Treatise on probability*. London: Macmillan.

Lagnado, D.A. & Shanks, D.R. (2002). Probability judgment in hierarchical learning: A conflict between predictiveness and coherence. *Cognition*, *83*, 81-112.

Levi, I. (2004). Jaakko Hintikka. *Synthese*, *140*, 37-41.

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgment. *Cognitive Science*, *26*, 181-206.

MacDonald, R.R. & Gilhooly, K.J. (1986). More about Linda, or conjunctions in context. *European Journal of Cognitive Psychology*, *2*, 57-70.

Massaro, D.W. (1994), A pattern recognition account of decision making. *Memory & Cognition*, *22*, 616-627.

Mellers, A., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269-275.

Milne, P. (1996). Log[$p(h/eb)/p(h/b)$] is the one true measure of confirmation. *Philosophy of Science*, *63*, 21–26.

Morier, D.M. & Borgida, E. (1984). The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychology Bulletin*, *10*, 243-252.

Nozick, R. (1981). *Philosophical Explanations*. Oxford (UK): Clarendon Press.

Olsson, E.J. (2005). Review of Bovens, L. & Hartmann, S., *Bayesian Epistemology*. *Studia Logica*, *81*, 289-292.

Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A. & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185-200.

Politzer, G. & Noveck, I.A. (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research*, *20*, 83-103.

Popper, K.R. (1954). Degree of confirmation. *British Journal for the Philosophy of Science*, *5*, 143-149.

Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, *30*, 191-198.

Sloman, S.A. & Lagnado, D. (2005). The problem of induction. In R. Morrison & K. Holyoak (eds.), *Cambridge handbook of thinking & reasoning* (pp. 95-116). New York: Cambridge University Press.

Sloman, S.A., Over, D., Slovak, L., & Stibel, J.M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, *91*, 296-309.

Slovic, P. & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*, 649-744.

Stich, S. (1990). *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation*. Cambridge, MA: MIT Press.

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science, 28*, 467-477.

Tentori, K., Crupi, V., & Osherson, D. (2007). Determinants of confirmation. *Psychonomic Bulletin & Review* (forthcoming).

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition, 103*, 107-119.

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

Tversky, A. & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. In T. Gilovich, D. Griffin, & D. Kahneman (eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 19-48). New York: Cambridge University Press.

Tversky, A. & Koehler, D.J. (1994). Support theory: A non-extensional representation of subjective probability. In T. Gilovich, D. Griffin, & D. Kahneman (eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 441-473). New York: Cambridge University Press.

Acknowledgements. Research supported by PRIN 2005 grant *Le dinamiche della conoscenza nella società dell'informazione* and by a grant from the SMC/Fondazione Cassa di Risparmio di Trento e Rovereto. Versions of this work have been presented at the Cognitive Psychology Laboratory, CNRS – University of Aix-Marseille I, at the workshop *Bayesianism, fundamentally*, Center for Philosophy of Science, University of Pittsburgh and at the conference *Confirmation, induction and science*, London School of Economics.

Notes

1. We are indebted to Nigel Harvey, Ralph Hertwig, David Lagnado and an anonymous referee for several important remarks on previous versions of this work.
2. The *conditional* confirmation condition (ii*) $c(h_2,e|h_1) > 0$ is equivalent, in probabilistic terms, to $p(e|h_1\&h_2) > p(e|h_1)$. The proof provided in the Appendix exploits that fact that the antecedent of the Theorem implies precisely $p(e|h_1\&h_2) > p(e|h_1)$ along with $p(e|\neg(h_1\&h_2)) < p(e|\neg h_1)$, which in turn imply $c(h_1\&h_2,e) > c(h_1,e)$. The latter implication is an instantiation of the so-called "weak law of likelihood", which holds for any Bayesian confirmation measure *c*, as already noticed by Joyce (2004) and Fitelson (2007).
3. Averaging models of the conjunction fallacy have been successfully tested by Fantino *et al.* (1997). Their results are consistent with the hypothesis proposed here, on the assumption that probability ratings reflect intuitive assessments of confirmation.
4. According to Ralph Hertwig, commenting on this paper, Hertwig & Gigerenzer's (1999) demonstration that "frequency" is more often given a mathematical interpretation as compared to "probability" does count as an explanation for the greater occurrence of violations of the conjunction rule under a probability framing. In our view it does not to the extent that the properties of the "non-mathematical" attributes which allegedly drive participants' probability judgments are left unspecified.
5. One may challenge the identification of representativeness with inverse probability and argue that, for better or worse, the concept of representativeness is fuzzy enough to accommodate the medical and Borg example. According to this reading, our own proposal could be seen as a sharpened account of the fit between evidence and hypothesis which representativeness was originally meant to capture. However, we tend to see the fuzziness of many uses of "representativeness" as a major limitation of this concept undermining its explanatory scope. By contrast, the theoretical grounds of the notion of confirmation are explicit, precise and open to thorough critical examination. (We thank an anonymous referee for raising this point.)

Appendix

Theorem. For any Bayesian measure of confirmation c among D, R, L, C, S and Z,

if (i) $c(h_1,e) \leq 0$
and (ii) $c(h_2,e|h_1) > 0$,
then $c(h_1\&h_2,e) > c(h_1,e)$

Proof:

We will prove the theorem by means of the following lemma:

Lemma. If $c(h_1,e) \leq 0$ and $c(h_2,e|h_1) > 0$, then:

(1) $p(e|h_1\&h_2) > p(e|h_1)$
(2) $p(e|\neg(h_1\&h_2)) < p(e|\neg h_1)$

Proof. (1) $c(h_2,e|h_1) > 0$ iff $c(e,h_2|h_1) > 0$ iff $p(e|h_1\&h_2) > p(e|h_1)$.
(2) $c(h_2,e|h_1) > 0$ iff $c(e,h_2|h_1) > 0$ iff $c(e,\neg h_2|h_1) < 0$ iff $p(e|\neg h_2\&h_1) < p(e|h_1)$. Since $c(h_1,e) \leq 0$, we have $p(e|h_1) \leq p(e|\neg h_1)$. Then it follows that $p(e|\neg h_2\&h_1) < p(e|\neg h_1)$, which is logically equivalent to $p(e|\neg(h_1\&h_2)) < p(e|\neg h_1)$.

By the lemma above, we will now prove the theorem considering measures D, R, L, C, S, and Z in turn. Notice that, since it is assumed that $c(h_1,e) \leq 0$, it is sufficient to prove the theorem in case $c(h_1\&h_2,e) \leq 0$ (for otherwise it would hold trivially).

Measure D:

$p(e|h_1\&h_2) > p(e|h_1)$ iff
$p(e|h_1\&h_2)/p(e) > p(e|h_1)/p(e)$ iff
$p(h_1\&h_2|e)/p(h_1\&h_2) > p(h_1|e)/p(h_1)$ iff
$[p(h_1\&h_2|e)/p(h_1\&h_2)] - 1 > [p(h_1|e)/p(h_1)] - 1$ iff
$[p(h_1\&h_2|e) - p(h_1\&h_2)]/p(h_1\&h_2) > [p(h_1|e) - p(h_1)]/p(h_1)$ iff
$[p(h_1\&h_2|e) - p(h_1\&h_2)] \times p(h_1) > [p(h_1|e) - p(h_1)] \times p(h_1\&h_2)$, which implies
$p(h_1\&h_2|e) - p(h_1\&h_2) > p(h_1|e) - p(h_1)$, i.e.,
$D(h_1\&h_2,e) > D(h_1,e)$

Measure R:

$p(e|h_1\&h_2) > p(e|h_1)$ iff
$p(e|h_1\&h_2)/p(e) > p(e|h_1)/p(e)$ iff
$p(h_1\&h_2|e)/p(h_1\&h_2) > p(h_1|e)/p(h_1)$ iff
$\ln[p(h_1\&h_2|e)/p(h_1\&h_2)] > \ln[p(h_1|e)/p(h_1)]$, i.e.,
$R(h_1\&h_2,e) > R(h_1,e)$

Measure L:

If $p(e|h_1\&h_2) > p(e|h_1)$ and $p(e|\neg(h_1\&h_2)) < p(e|\neg h_1)$, then
$p(e|h_1\&h_2)/p(e|\neg(h_1\&h_2)) > p(e|h_1)/p(e|\neg h_1)$, which implies
$\ln[p(e|h_1\&h_2)/p(e|\neg(h_1\&h_2))] > \ln[p(e|h_1)/p(e|\neg h_1)]$, i.e.,
$L(h_1\&h_2,e) > L(h_1,e)$

Measure C:

$D(h_1\&h_2,e) > D(h_1,e)$ iff
$D(h_1\&h_2,e) \times p(e) > D(h_1,e) \times p(e)$, i.e.,
$C(h_1\&h_2,e) > C(h_1,e)$

Measure S:

$D(h_1\&h_2,e) > D(h_1,e)$ iff
$D(h_1\&h_2,e)/p(\neg e) > D(h_1,e)/p(\neg e)$, i.e.,
$S(h_1\&h_2,e) > S(h_1,e)$

Measure Z:

$p(e|h_1\&h_2) > p(e|h_1)$ iff
$p(e|h_1\&h_2)/p(e) > p(e|h_1)/p(e)$ iff
$p(h_1\&h_2|e)/p(h_1\&h_2) > p(h_1|e)/p(h_1)$ iff
$[p(h_1\&h_2|e)/p(h_1\&h_2)] - 1 > [p(h_1|e)/p(h_1)] - 1$ iff
$[p(h_1\&h_2|e) - p(h_1\&h_2)]/p(h_1\&h_2) > [p(h_1|e) - p(h_1)]/p(h_1)$, i.e.,
$Z(h_1\&h_2,e) > Z(h_1,e)$