# Probability, Confirmation, and the “Conjunction Fallacy”

**Branden Fitelson**¹  
Department of Philosophy & Religion  
Northeastern University  
branden@fitelson.org  
http://fitelson.org/

---

¹ This is based on joint work with Vincenzo Crupi (Turin) & Katya Tentori (Trento).

---

### Overview

- **Some Background on Probability (Blackboard)**
- **Historical Background I: Carnapian Confirmation Theory**
  - Confirmation, “Logical” Probability, and Relevance
- **Historical Background II: Bayesian Confirmation Theory**
  - Confirmation and Subjective Probabilistic Relevance
- **Applying Bayesian Confirmation to the “Conjunction Fallacy”**
  - The Traditional “Conjunction Fallacy” Cases
  - A Confirmation-Theoretic Approach to the Traditional CFs
  - Non-Traditional CFs and Bayesian Confirmation Theory

---

In the first edition of *LFP*, Carnap [3] undertakes a precise probabilistic explication of the concept of confirmation. This is where modern confirmation theory was born (in sin).

Carnap was interested mainly in *quantitative* confirmation (which he took to be fundamental). But, he also gave (derivative) qualitative and comparative explications:

- **Qualitative.** $E$ inductively supports $H$.
- **Comparative.** $E$ supports $H$ more strongly than $E'$ supports $H'$.
- **Quantitative.** $E$ inductively supports $H$ to degree $r$.

Carnap begins by clarifying the *explicandum* (the informal “inductive support” concept) in various ways, including:

- **Qualitative.** ($\star$) $E$ gives some (positive) evidence for $H$.
- **Comparative.** $E$ supports $H$ more strongly than $E'$ supports $H'$.
- **Quantitative.** $E$ inductively supports $H$ to degree $r$.

Note two things. First, ($\star$) sounds *epistemic* (not *logical*). Second, ($\star$) sounds like it involves (positive) *relevance*.

Strangely, Carnap proceeds (in *LFP$_1$*) to offer a *logical* account of confirmation that does *not* involve relevance.

These were the two original sins of Bayesian confirmation…
- In the 1st ed. of LFP, Carnap characterizes “the degree to which \( E \) confirms \( H \)” as \( c(H, E) = \Pr(H \mid E) \), which leads to:
  - **Quantitative.** \( \Pr(H \mid E) = r \).
  - **Comparative.** \( \Pr(H \mid E) > \Pr(H' \mid E') \).
  - **Qualitative.** \( \Pr(H \mid E) > t \) (typically, with “threshold” \( t > \frac{1}{2} \)).
  - Doesn’t sound like (\( * \)). More on this dissonance below.
- Like Hempel [8], Carnap wanted a logical explication of confirmation (as a relation between sentences in FOLs).
- For Carnap, this meant that the probability functions used in confirmation theory must themselves be logical.
- This leads naturally to the Carnapian project of providing a “logical explication” of conditional probability \( \Pr(\cdot \mid \cdot) \) itself.
- Here, Carnap was strongly influenced by Keynes [10], who believed there were (probabilistic) “partial entailments”. I’m somewhat skeptical [6] (as are most modern Bayesians).
- Hempel’s theory of confirmation [8] satisfies the following: (SCC) If \( E \) confirms \( H \), then \( E \) confirms all consequences of \( H \).

\[ c_\text{i}(H, E) = \Pr(H \mid E) \]

- \( c_\text{i} \) is more similar to (\( * \)) than \( c_\text{f} \) is. To see this, note that we can have \( \Pr(H \mid E) > t \) even if \( E \) lowers the probability of \( H \).
- Example: Let \( H \) be the hypothesis that John does not have HIV, and let \( E \) be a positive test result for HIV from a highly reliable test. Plausibly, in such cases, we could have both:
  - \( \Pr(H \mid E) > t \), for just about any threshold value \( t \), but
  - \( \Pr(H \mid E) < \Pr(H) \), since \( E \) lowers the probability of \( H \).
- So, if we adopt Carnap’s \( c_\text{f} \)-explication, then we must say that \( E \) confirms \( H \) in such cases. But, in (\( * \))-terms, this implies \( E \) provides some positive evidential support for \( H \)!
- I take it we don’t want to say that. Intuitively, what we want to say here is that, while \( H \) is (still) highly probable given \( E \), (nonetheless) \( E \) provides (strong!) evidence against \( H \).
- Carnap [2] seems to appreciate this dissonance, when he concedes \( c_\text{i} \) is (in some settings) “more interesting” than \( c_\text{f} \).
- Contemporary Bayesians would agree with this. They’ve since embraced a probabilistic relevance conception [13].
Bayesianism is based on the assumption that the degrees of belief (or credences) of rational agents are **probabilities**.

Let \( \Pr(H) \) be the degree of belief that a rational agent \( a \) assigns to \( H \) at some time \( t \) (call this \( a \)'s “prior” for \( H \)).

Let \( \Pr(H \mid E) \) be the degree of belief that \( a \) would assign to \( H \) (just after \( t \)) were \( a \) to learn \( E \) at \( t \) (\( a \)'s "posterior" for \( H \)).

Toy Example: Let \( H \) be the proposition that a card sampled from some deck is a ♠, and \( E \) assert that the card is black.

Making the standard assumptions about sampling from 52-card decks, \( \Pr(H) = \frac{1}{4} \) and \( \Pr(H \mid E) = \frac{1}{2} \). So, learning that \( E \) **raises the probability** one (rationally) assigns to \( H \).

Following Popper [12], Bayesians define confirmation in a way that is **formally** very similar to Carnap’s \( c_i \)-explication.

For Bayesians, \( E \) confirms \( H \) for an agent \( a \) at a time \( t \) iff \( \Pr(H \mid E) > \Pr(H) \), where \( Pr \) captures \( a \)'s credences at \( t \).

While this is **formally** very similar to Carnap’s \( c_i \), it uses credences as opposed to “logical” probabilities [13], [6].

When it comes to **quantitative** judgments, Bayesians use various **relevance measures** \( c \) of degree of confirmation.

These are much like the candidate functions \( f \) we saw in connection with Carnapian \( c_i \), but defined relative to subjective probabilities rather than “logical” probabilities.

There are **many comparatively distinct** measures. See [5] and [17] for philosophical and psychological discussion.

Once we choose a measure \( c(H, E) \) of the degree to which \( E \) confirms \( H \), we can explicate **comparative** confirmation relations. E.g., \( E \) favors \( H_1 \) over \( H_2 \) iff \( \Pr(H_1, E) > \Pr(H_2, E) \).

Note: \( \Pr(H \mid E) \) is a **bad** candidate for \( c(H, E) \) in this context. It implies “\( E \) favors \( H_1 \) over \( H_2 \),” in some cases where \( E \) is negatively relevant to \( H_1 \) but positively relevant to \( H_2 \) [12]!

In the context of **comparative** confirmation, there is ongoing philosophical/theoretical debate about the appropriate choice of \( c \) (e.g., the Likelihoodism debate [7]).

An account is **robust** if it does not depend on choice of \( c \).

Tversky and Kahneman [19] discuss the following example, which was the first example of the “conjunction fallacy”:

\( (E) \) Linda is 31, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and she also participated in antinuclear demonstrations.

Is it more probable, given \( E \), that Linda is \( (H_1) \) a bank teller, or \( (H_1) \) and \( H_2 \) a bank teller and an active feminist?

Most say \( “H_1 \) and \( H_2” \) is more probable (given \( E \)) than \( H_1 \). On its face, this violates comparative probability theory, since \( X \equiv Y \) implies \( \Pr(X \mid E) \leq \Pr(Y \mid E) \), and \( H_1 \) and \( H_2 \equiv H_1 \).

Experiments have been done to ensure subjects understand “\( H_1 \) and \( H_2 \)” in the experiment as a **conjunction** \( H_1 \) and \( H_2 \), and \( H_1 \) as a **conject** thereof (not as \( H_1 \) & \( \sim H_2 \) [15, 16].

At the same time, the “fallacy” persists when people are queried about **betting odds** rather than **probabilities** [15, 1].

Comparative Bayesian confirmation can be helpful [11]. We’re developing detailed accounts along these lines [4].
The first inequality (i) has already been empirically well established in several traditional (Linda-like) CF cases [14].

- Our (ii)/(ii*) have not been explicitly tested. But, we suspect these will obtain (empirically) in the the traditional CF cases.

- We are performing experiments to test the (i)/(ii) and (i)/(ii*) accounts of the traditional CF cases [4]. Preliminary results indicate that (ii) is commonly endorsed by subjects.

- Interestingly, many seem to judge (i) & (ii) as more plausible than (i) & (ii*) [ii] vs (ii*]). Do you? Note: (i) & (ii) = (ii*)!

- This suggests (i) & (ii) may provide a more robust explanation than (i) & (ii*) for traditional CFs (a meta-CF?).

- But, there are other (non-traditional) sorts of CF cases in which (ii)/(ii*) seem false, and (i) alone does not seem sufficient to predict all patterns of response (next slide).

- Even in this broader class of CFs, however, we think that some confirmation-theoretic conditions will be useful for predicting and explaining observed patterns of response.

- Here is a non-traditional CF example: 

- $E =$ John is Scandanavian; $H_1 =$ John has blue eyes; $H_2 =$ John has blond hair.

- In this case, (i) seems plausible, but (ii)/(ii*) do not.

- Moreover, it is not at all clear whether this is (normatively) a case in which we should have $c(H_1 \land H_2, E) > c(H_1, E)$.

- Descriptively, we suspect confirmation-theoretic relations between $H_1$ and $H_2$ themselves may be involved in the CF.

- Specifically, the terms $c(H_1, H_1)$ seem to be salient. We bet they are explanatorily relevant. There is some preliminary evidence which supports this conjecture [18].

- Psychologically, we think there are two important sets of confirmation-theoretic factors involved in CF cases:

- $c(H_1, E), c(H_2, E), c(H_1 \land H_2), c(H_2, E \mid H_1)$. [Traditional CF]

- $c(H_1, H_2), c(H_2, H_1), c(H_1, H_2 \mid E), c(H_2, H_1 \mid E)$. [NT CF]

- More general confirmaiton-theoretic models have recently been developed which seem to subsume and explain all known instances of the “conjunction fallacy” [18].