Explaining the Conjunction Fallacy

Borut Trpin (borut.trpin@lrz.uni-muenchen.de)
Munich Center for Mathematical Philosophy, LMU Munich, 80539 Munich (Germany)

Stephan Hartmann (S.Hartmann@lmu.de)
Munich Center for Mathematical Philosophy and Department of Psychology, LMU Munich, 80539 Munich (Germany)

Abstract

The conjunction fallacy (CF) describes a pattern where individuals disregard the principles of probability by assessing certain conjunctive statements as more probable than the individual parts of those statements. The fallacy may be fruitfully reconstructed as the normatively correct assessment of something else than probability, for instance of inductive confirmation or coherence. We argue that these approaches have some counter-intuitive consequences in scenarios that have not yet been experimentally tested. We then suggest a novel explanation of the CF according to which the fallacious reasoning arises due to an assessment of explanatory power.

Keywords: Conjunction Fallacy; Probabilistic Reasoning; Explanation; Confirmation; Coherence; Formal Epistemology

Introduction

The conjunction fallacy (CF) is one of the most discussed examples of fallacious reasoning in cognitive science (see, e.g., Tversky & Kahneman, 1982, 1983; Morier & Borgida, 1984; Fiedler, 1988; Hertwig & Gigerenzer, 1999; Sides, Osherson, Bonini, & Viale, 2002; Tentori, Bonini, & Osherson, 2004; Crupi, Fitelson, & Tentori, 2008; Crupi, Elia, Aprà, & Tentori, 2018; Jönsson & Shogenji, 2019). The fallacy occurs when a conjunction of multiple propositions is considered to be more probable than a single conjunct. This is surprising as it is a very elementary law of probability theory that a conjunction of multiple information items cannot be more probable than any of the individual information items that make up the conjunction. Formally,

$$P(\mathrm{H}_1, \mathrm{H}_2) \leq P(\mathrm{H}_i), \quad i \in \{1, 2\}.$$

Yet, the CF is a very robust and highly replicated phenomenon, which can easily be elicited if the conjunction intuitively makes more sense than the individual information on its own. Consider the well-known Linda problem from Tversky and Kahneman (1982, p. 92):

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Rank the following statements by their probability:[1]
>
> 1. Linda is a bank teller.
> 2. Linda is a bank teller and active in the feminist movement.

The vast majority of experimental participants systematically rank the conjunction "Linda is a bank teller and active in the feminist movement" as more probable than "Linda is a bank teller" even though it is easy to see that every feminist bank teller is also a bank teller, hence the more specific conjunction cannot be more probable.

Several models acknowledge the CF as a genuine fallacy and seek to explain why it occurs by referring, e.g., to the representativeness heuristic (Tversky & Kahneman, 1982, 1983) or confirmation- and coherence-based reasoning (Crupi et al., 2008; Jönsson & Shogenji, 2019). Other models deny the existence of any fallacy and attribute participants' responses to a misunderstanding between the experimenter and the participants, e.g. due to conversational implicatures (Dulany & Hilton, 1991) or the wrong format of probability (Gigerenzer, 1991; Hertwig & Gigerenzer, 1999). It has also been noted that no fallacy occurs if the assessment of the reliability of the respective information sources is explicitly factored in (Bovens & Hartmann, 2003; Hartmann & Meijs, 2012).

Our contribution belongs to the former approach as we acknowledge the CF as a genuine fallacy. We, however, seek to explain the CF by referring to explanatory considerations. Specifically, we will lay out the theoretical foundations for a novel Bayesian model according to which the CF is expected when a conjunction of options provides a better explanation of the setup (or the other way around). Using a precise probabilistic measure of explanatory power also affords us to make predictions about how expected the CF is in some case – the greater the explanatory power of the conjunction compared to a non-conjunction, the more expected the CF.

Moreover, we will argue that while both the confirmation- and coherence-based models point in the right direction, they are just special cases of explanatory reasoning as these models reduce the situation to a two-place assessment of the options (that are ranked) and the setup (i.e. the background story or vignette). Our approach provides an account of the fal-

[1] Most experiments of the CF include other statements in the ranking task, often also the conjunction with an explicit negation of the plausible conjunct and several filler statements (cf. Tentori et al., 2004). Here, we only include the target statements (the conjunction and the individual conjunct) for brevity.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

lacy which can explain several variations and, because it does not reduce the situations to a two-place relationship, provides some interesting novel predictions that may be tested in future empirical work. The present contribution may therefore be seen as a normatively grounded attempt at modeling the CF, which can then be extended to more complex situations.

Confirmation- and Coherence-Based Accounts

An influential account of the CF due to Tentori, Crupi and their colleagues (Crupi et al., 2008; Tentori & Crupi, 2012; Tentori, Crupi, & Russo, 2013; Crupi et al., 2018; Cevolani & Crupi, 2022) proposes that the fallacious inference arises due to an intuitive assessment of confirmatory imports. The CF then depends on whether the additional conjunct increases the confirmation of the individual conjunct. The greater the increase in confirmation, the greater the expectation of the CF. According to this account, people act as intuitive philosophers of science when assessing the CF-scenarios:[2] similarly as in scientific research, people assess whether evidence supports (i.e., confirms) a hypothesis and do not, at least not in the first place, assess the probability of the hypothesis as such.

This idea may be easily formalized. In probabilistic terms, evidence E confirms a hypothesis H if and only if $P(\mathrm{H}|\mathrm{E}) - P(\mathrm{H}) > 0$, and disconfirms it if the left-hand side is negative. Two simplest measures of confirmation of H by E are then defined as

$$\mathcal{C}_d(\mathrm{H};\mathrm{E}) = P(\mathrm{H}|\mathrm{E}) - P(\mathrm{H})$$
$$\text{and } \mathcal{C}_r(\mathrm{H};\mathrm{E}) = \frac{P(\mathrm{H}|\mathrm{E})}{P(\mathrm{H})}.$$

(See Fitelson, 1999 for further measures and a discussion of their merits and limitations.) Importantly, a conjunctive hypothesis may be better confirmed than an individual conjunct. That is, it can be the case that $\mathcal{C}(\mathrm{H}_1,\mathrm{H}_2;\mathrm{E}) > \mathcal{C}(\mathrm{H}_1;\mathrm{E})$.[3] In Linda's case, for example, Crupi et al. (2008) argue that her description should be taken as evidence E, Linda being a bank teller as H_1, and Linda being a feminist as H_2. Clearly, $P(\mathrm{H}_1|\mathrm{E})$ is either similar or less than $P(\mathrm{H}_1)$: the description reduces the probability of Linda being a bank teller. On the other hand, the conjunction is more probable in light of the description than on its own: $P(\mathrm{H}_1,\mathrm{H}_2|\mathrm{E}) > P(\mathrm{H}_1,\mathrm{H}_2)$.

The authors then provide a more general approach to several variants of the CF by considering three determinants: How much the individual hypothesis H_1 is confirmed by the story, how much H_2 confirms H_1 in light of the background story E, and how much E confirms the story in light of H_2. Hence, they model the expectation of the conjunction fallacy as (see Tentori et al., 2013, Eq. 13 on p. 248):

1. a decreasing function of $\mathcal{C}(\mathrm{H}_1;\mathrm{E})$,
2. an increasing function of $\mathcal{C}(\mathrm{H}_1;\mathrm{E}|\mathrm{H}_2)$, and
3. an increasing function of $\mathcal{C}(\mathrm{H}_1;\mathrm{H}_2|\mathrm{E})$.

This is important as there are several variants of the CF. Tversky and Kahneman (1983), for instance, make a distinction between the so-called M-A and A-B cases of the CF. The former, M-A, refers to the scenarios such as Linda's where the description ("Model") is positively related to one target statement (statement "A") and negatively related to another (statement "B"). The conjunction of A and B is then typically considered as more probable than B. The A-B cases, on the other hand, represent the scenarios where the "Model" is negatively or not at all related to the target statement B, but another statement A stands in a positive relation to B (see Figure 1).

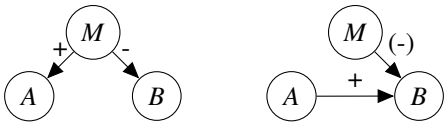

Figure 1: Two basic types of the conjunction fallacy adapted from Tversky and Kahneman (1983): M-A (left) and A-B (right). The arrows represent correlations.

Consider the following A-B case (Tversky & Kahneman, 1983, p. 305):

> A health survey was conducted in a representative sample of adult males in British Columbia of all ages and occupations. Mr. F. was included in the sample. He was selected by chance from the list of participants. Which of the following statements is more probable? (check one)
>
> 1. Mr. F. has had one or more heart attacks.
> 2. Mr. F. has had one or more heart attacks and he is over 55 years old.

Experimental participants reliably rank the conjunction as more probable in this and similar cases. This would be hard to explain in confirmation-based terms if we used the same strategy as in the M-A cases. The background story is probabilistically independent from the statements: Mr. F. was selected by chance, so $P(\mathrm{H}_1|\mathrm{E}) = P(\mathrm{H}_1)$ and $P(\mathrm{H}_1,\mathrm{H}_2|\mathrm{E}) = P(\mathrm{H}_1,\mathrm{H}_2)$, which clearly means that the story has no confirmatory import. However, because the confirmation-based approach also considers how much H_2 confirms H_1, the case becomes non-problematic. An age of over 55 clearly increases the probability of a heart attack, so $\mathcal{C}(\mathrm{H}_1;\mathrm{H}_2|\mathrm{E})$ is higher than $\mathcal{C}(\mathrm{H}_1;\mathrm{E})$. The only potential issue is that the two determinants, $\mathcal{C}(\mathrm{H}_1;\mathrm{H}_2|\mathrm{E})$ and $\mathcal{C}(\mathrm{H}_1;\mathrm{E}|\mathrm{H}_2)$ may pull in different directions. For instance, in the M-A scenarios, H_1 and H_2 often happen to be at odds and disconfirm one another.

Jönsson and Shogenji (2019) argue that we should therefore instead assess the coherence of the statements and the

[2] Interestingly, a similar analogy of everyday reasoners as naïve philosophers of science also plays a role in an unrelated but very influential Bayesian reconstruction of another well-known reasoning fallacy, the Wason selection task (Wason, 1968) by Oaksford and Chater (1994).

[3] We use $\mathcal{C}(\mathrm{H};\mathrm{E})$ to refer to confirmation in the most general sense of $P(\mathrm{H}|\mathrm{E})$ being greater than $P(\mathrm{H})$, without referring to any specific measure of confirmation. This is because all measures of confirmation conform to this general principle.

description. There are several ways how to probabilistically measure the coherence of an information set (see Olsson, 2022 for a review). Jönsson and Shogenji (2019) first consider the measure proposed by Shogenji (1999):

$$coh_{Sh}(\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_n\}) = \frac{P(\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_n)}{P(\mathrm{E})\cdot P(\mathrm{H}_1)\cdots P(\mathrm{H}_n)}$$

Due to several issues with that measure they then suggest a revised measure:

$$coh_{Sh+}(\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_n\}) = coh_{Sh}(\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_n\})^{\frac{1}{n-1}}$$

The revised measure has the interesting property that "[t]o raise the degree of confirmation the added member must be more coherent with the conjunction of the extant members than the extant members are among themselves" (Jönsson & Shogenji, 2019, p. 234). Formally, (for the derivation, see ibid.):

$$\begin{aligned} & coh_{Sh+}(\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_{n+1}\}) > coh_{Sh+}(\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_n\}) \\ \iff \quad & coh_{Sh+}(\{\mathrm{E}\wedge\mathrm{H}_1\wedge\cdots\wedge\mathrm{H}_n,\mathrm{H}_{n+1}\}) \\ & \qquad > coh_{Sh+}(\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_n\}) \end{aligned}$$

The CF typically consists of three elements, the story, E, and two conjuncts, H_1 and H_2. According to the approach advocated by Jönsson and Shogenji (2019), the CF is then expected when $coh_{Sh+}(\{\mathrm{E},\mathrm{H}_1,\mathrm{H}_2\}) > coh_{Sh+}(\{\mathrm{E},\mathrm{H}_1\})$. Note that if the set $\mathbf{S}$ is an information pair ($n = 2$; e.g., $\mathbf{S} = \{\mathrm{E},\mathrm{H}\}$), then $coh_{Sh+}(\mathbf{S}) = coh_{Sh}(\mathbf{S})$. Combining this observation and the mentioned formal property, the CF is then expected when

$$\begin{aligned} & coh_{Sh+}(\{\mathrm{E},\mathrm{H}_1,\mathrm{H}_2\}) > coh_{Sh+}(\{\mathrm{E},\mathrm{H}_1\}) \\ \iff & coh_{Sh+}(\{\mathrm{E}\wedge\mathrm{H}_1,\mathrm{H}_2\}) > coh_{Sh+}(\{\mathrm{E},\mathrm{H}_1\}) \\ \iff & coh_{Sh}(\{\mathrm{E}\wedge\mathrm{H}_1,\mathrm{H}_2\}) > coh_{Sh}(\{\mathrm{E},\mathrm{H}_1\}) \\ \iff & \frac{P(\mathrm{H}_1\wedge\mathrm{E},\mathrm{H}_2)}{P(\mathrm{H}_1\wedge\mathrm{E})P(\mathrm{H}_2)} > \frac{P(\mathrm{H}_1,\mathrm{E})}{P(\mathrm{H}_1)P(\mathrm{E})} \\ \iff & \frac{P(\mathrm{H}_1\wedge\mathrm{E}|\mathrm{H}_2)}{P(\mathrm{H}_1\wedge\mathrm{E})} > \frac{P(\mathrm{H}_1|\mathrm{E})}{P(\mathrm{H}_1)} \\ \iff & \mathcal{C}_r(\mathrm{H}_1\wedge\mathrm{E};\mathrm{H}_2) > \mathcal{C}_r(\mathrm{H}_1;\mathrm{E}). \end{aligned}$$

In Linda's case this means that the conjunction fallacy is expected because Linda being a feminist, H_2, confirms (i.e., raises the probability of) the conjunction of Linda being a bank teller and her description, $\mathrm{H}_1\wedge\mathrm{E}$, more than Linda's description, E, confirms her being a bank-teller, H_1.

This is an interesting insight: the coherence-based account of the CF is actually also confirmation-based. The only difference to the approach by Tentori, Crupi and their colleagues is that Jönsson and Shogenji (2019) consider how much the additional conjunct confirms the rest of the scenario. Unfortunately, all confirmation-based accounts have a limited scope of application because they assess a two-place relation between two statements, the one that does the confirming and the one that is being confirmed. Consider the following (novel) example to see why this may be an issue:

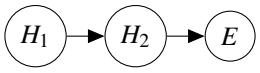

Figure 2: A probabilistic chain network which corresponds to the scenario of a fertilized (H_1) and brooded upon (H_2) egg that hatched (E). Note the distinction between binary propositional variables, e.g. H_1 (in italic script) and their values, e.g. H_1 and $\neg\mathrm{H}_1$ (in roman script).

> An egg hatched into a chick (E). You are not provided with any additional details about the eggs or the chicks. Which hypothesis is more probable?
>
> - H_2: A hen was brooding on the egg.
> - $\mathrm{H}_1,\mathrm{H}_2$: The egg was fertilized and a hen was brooding on it.

We predict that most individuals would consider the conjunctive hypothesis $\mathrm{H}_1,\mathrm{H}_2$ as more probable than hypothesis H_2 individually. This predicted ranking arises from the intuition that hatching is better explained by a combination of fertilization and brooding than by brooding on its own. However, because this example corresponds to a positively correlated probabilistic chain going from H_1 through H_2 to E (see Figure 2), the confirmation-based accounts fail.

Let us now consider this example in the confirmation-based approach by Tentori, Crupi and colleagues. On their account, the CF is not expected here: (1) H_2 (brooding) is confirmed by E (chick), which reduces the expectancy of the CF; (2) H_1 (fertilization) is neither confirmed nor disconfirmed by E given H_2 as H_2 screens off H_1 from E, so this determinant is irrelevant for the CF; (3) H_1 (fertilization) is at most mildly (if at all) confirmed by H_2 (brooding) given E (chick), which also is not favorable for the CF.

The coherence-based approach by Jönsson and Shogenji (2019) also gives a similar counter-intuitive prediction. Suppose that we explicitly mention some intuitively plausible aspects of the situation – that a hen is very likely to brood a fertilized egg (e.g., $P(\mathrm{H}_2|\mathrm{H}_1) = .8$) and unlikely to brood a non-fertilized egg (e.g., $P(\mathrm{H}_2|\neg\mathrm{H}_1) = .2$). Similarly, we may mention that the egg is very likely to hatch if it is brooded on (e.g., $P(\mathrm{E}|\mathrm{H}_2) = .8$) and very unlikely to hatch if it is not brooded on (e.g., $P(\mathrm{E}|\neg\mathrm{H}_2) = .05$). In such a scenario, the coherence-based account predicts that the conjunction of brooding and fertilization should never be preferred over the brooding hypothesis on its own, regardless of how probable it is that the egg is fertilized in the first place. See Figure 3.

Our aim in this contribution is to lay out the theoretical foundations of a new approach, so we use the example with the chick merely to illustrate the possibility of counterexamples to confirmation- and coherence-based accounts of the CF. An example that would be used in an actual experiment may need to be adapted and it would also make sense to explicitly include the option $\neg\mathrm{H}_1,\mathrm{H}_2$: "The egg was not fertilized and a hen was brooding on it", as well as filler items. However, even if we go with different examples, the prob-

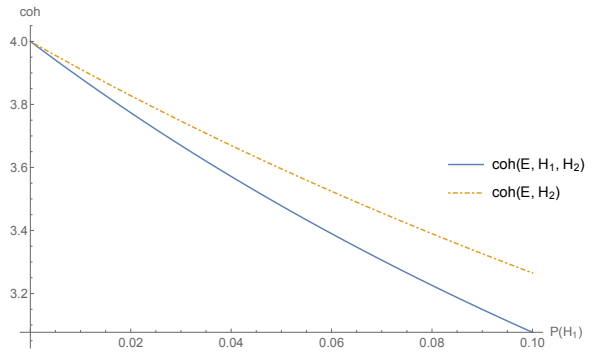

Figure 3: Coherence (measured by coh_{Sh+}) of the conjunctive hypothesis with the evidence of a hatched chick (solid blue line) and the coherence of the individual conjunct with the same evidence (dashed blue line) as a function of the probability that the egg was fertilized. As the conjunctive case is always less coherent, the account surprisingly predicts no CF.

lems in general arise when we deal with positively correlated probabilistic chains, i.e. those where H_2 confirms H_1 which in turn confirms E. In most cases like this, we expect the CF (i.e., that H_1, H_2 are judged as more probable than H_1), although the confirmation- and coherence-based accounts will predict otherwise.[4]

This suggests that the explanation of the conjunction fallacy must be found elsewhere than in confirmation- or coherence-based accounts. We propose that the CF instead arises due to explanatory reasoning. When considering the statements in question we are, on our proposed account, instead resolving the *why* question: Why did the evidence obtain? Do we have a reasonable explanation? For instance, when assessing the Linda case, we may be intuitively wondering how well Linda's description explains that Linda is a bank teller – not very well. Hence, Linda being a bank teller makes little sense to us. On the other hand, Linda's story better explains (i.e., makes more sense of) her becoming a feminist bank teller. And indeed, explanatory reasoning presents an important aspect of human reasoning (see, e.g., Lombrozo & Carey, 2006; Lombrozo, 2006, 2012). Let us therefore turn to this novel approach which combines the strengths of the confirmation- and coherence-based accounts of the conjunction fallacy and supplements them in scenarios where they fall short.

Measuring the Explanatory Power

The explanation-based approach to conjunction fallacy is rather simple: the main principle is that people commit the conjunction fallacy in either of the following two cases:

[4]We can use this strategy to generate countless other examples. As another example, suppose E: "The music score enhances the emotional experience of a movie". We assume that participants would assess the conjunction of H_1: "A composer wrote the music score for the movie" and H_2: "An orchestra recorded the music score" as more probable than H_2 individually in light of E. The confirmation- and coherence-based accounts predict the contrary.

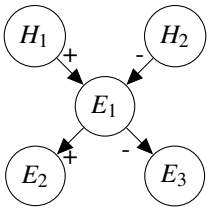

Figure 4: An example of a probabilistic network which could generate several different predictions regarding the conjunction fallacy. The +/- signs indicate positive or negative probabilistic relevance along the arc. In this case we may expect, assuming E_1 is the description, that any combination of + and - nodes will be ranked as more "probable" than the - node alone, e.g. $P(H_1, H_2) > P(H_2), P(H_1, E_3) > P(E_3)$ and so on.

1. When the explanation of evidence E by hypotheses H_1 and H_2 has greater explanatory power than that provided by H_1 or H_2 individually, or

2. When a hypothesis has greater explanatory power with respect to evidence E_1 and E_2 than with respect to evidence E_1 or E_2 individually.

The larger the difference in the explanatory power in these cases, the more likely the conjunction fallacy is to occur. Note that this opens the possibility also to yet unexplored cases of multiple conjunction fallacies in cases where two hypotheses H_1 and H_2 have greater explanatory power with respect to E_1 and E_2 than any other combination or other yet unexplored combinations (Figure 4 suggests a few options). However, before we can fully engage with these possibilities, we need to also have a way of measuring the degree of explanatory power. Fortunately, the philosophy of science literature comes with a multitude of the measures (Schupbach & Sprenger, 2011; Crupi & Tentori, 2012; Hartmann & Trpin, 2023a). Most of the measures are confirmation-based in the sense that they quantify how much H confirms E. Some also penalize for how much H confirms not-E (Cohen, 2016). As Roche and Sober (2023) convincingly show, these confirmation-based measures of explanatory power turn out to be unsatisfactory because they attempt to analyze rather complex scenarios by assessing a two-place relation (of the explanans and the explanandum) only.

Following Hartmann and Trpin (2023a) we therefore suggest that a coherence-based measure of explanatory power should be used instead. The idea of this approach is that the explanatory power of some hypothesis H_1 (or several hypotheses $H_1, \ldots, H_n$) with respect to evidence E should be determined in a contrastive way by "answering" why E and not rather $\neg$E. The explanatory power then depends on two determinants: (1) the coherence of the hypotheses and the evidence, $coh(\{H_1, \ldots, H_n, E\})$ and (2) the coherence of the hypotheses and the negation of the evidence: $coh(\{H_1, \ldots, H_n, \neg E\})$. The first part, therefore, corresponds to the coherence-based approach to the CF, which we have discussed already, while the second part goes beyond it.

As Hartmann and Trpin (2023a) show, it is important to respect two principles when measuring coherence: (1) how an information set is statistically correlated and (2) how much relative overlap there is among information items. The former may be obtained with the measure coh_{Sh} that we have already discussed, while the latter may be done by using coh_{OG} proposed by Olsson (2002) and Glass (2002):

$$coh_{OG}(\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_n\}) = \frac{P(\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_n)}{P(\mathrm{E}\vee\mathrm{H}_1\vee\mathrm{H}_n)}$$

A measure of coherence that combines both principles, coh_{OG+} is then defined as follows (Hartmann & Trpin, 2023b):

$$coh_{OG^+}(\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_\mathrm{n}\}) = \frac{coh_{OG}(\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_\mathrm{n}\})}{coh_{OG}(\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_\mathrm{n}\}^*)} = \frac{P(\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_n)}{P(\mathrm{E}\vee\mathrm{H}_1\vee\mathrm{H}_n)} \Big/ \frac{P(\mathrm{E})\cdot P(\mathrm{H}_1)\cdots P(\mathrm{H}_n)}{1-P(\neg\mathrm{E})\cdot P(\neg\mathrm{H}_1)\cdots P(\neg\mathrm{H}_n)}$$

where $\{\mathrm{E},\mathrm{H}_1,\ldots,\mathrm{H}_\mathrm{n}\}^*$ represents a set in which the propositions are assumed to be probabilistically independent of one another but the marginal probabilities of propositions remain unchanged. In non-formal terms: to assess the coherence of an information set, we need to consider how much the propositions actually overlap relative to how much they would overlap if they were probabilistically independent.

Plugging this measure into the described coherence-based approach to explanatory power then basically means that in assessing a set of potential evidence and hypotheses $\mathbf{S} = \{\mathrm{H}_1,\ldots,\mathrm{H}_n,\mathrm{E}_1,\ldots,\mathrm{E}_m\}$ we proceed as follows:

$$\mathcal{E}_{coh_{OG^+}}(\mathbf{S}) = \frac{coh_{OG^+}(\mathbf{S}) - coh_{OG^+}(\mathbf{S}_{\neg\mathbf{E}})}{coh_{OG^+}(\mathbf{S}) + coh_{OG^+}(\mathbf{S}_{\neg\mathbf{E}})}$$

where $\mathbf{S}_{\neg\mathbf{E}} = \{\mathrm{H}_1,\ldots,\mathrm{H}_n,\neg\mathrm{E}_1,\ldots,\neg\mathrm{E}_m\}$ is the corresponding set where the evidential information is negated. The approach therefore considers how coherent hypotheses and the evidence are and then sets this off against the case in which the evidence is negated.

We can then predict that the CF will occur when $\mathcal{E}_{coh_{OG^+}}(\mathbf{S_{n+1}}) > \mathcal{E}_{coh_{OG^+}}(\mathbf{S_n})$, where n refers to the number of elements involved in the model. In other words, if including an additional hypothesis or piece of evidence increases the explanatory power of the hypotheses over the evidence, then the CF is expected. The greater the explanatory impact of the additional information, the greater the likelihood of the CF.

By following this approach we can then show that the standard cases may be resolved:

The M-A paradigm (the Linda-like cases): The probabilistic dependencies in these cases may be reconstructed by a common cause network $E_1 \leftarrow H \rightarrow E_2$, where H stands for the hypothetical description or context, and E_1 and E_2 correspond to the binary variables of which one is likely and the other one is not in light of H. Suppose E_1 is negatively correlated to H. Then it is easy to see that $\mathcal{E}_{coh_{OG^+}}(\{\mathrm{H},\mathrm{E}_1\}) < 0$ because $coh_{OG^+} > 1$ for positively correlated sets of 2 or 3 information items (and just the contrary for negatively correlated sets of 2 or 3 items; for a proof see Hartmann & Trpin, 2023b). H is therefore bad in explaining E_1. The question then is whether and how adding the positively correlated E_2 affects the explanatory power of H.

Let us show how this may play out in Linda's case. We assume a common cause network $E_1 \leftarrow H \rightarrow E_2$, where the binary variable H stands for Linda's description (which may be true or false), and E_1 and E_2 respectively stand for Linda being a bank-teller and Linda being a feminist. Furthermore, let us assume that $P(\mathrm{E}_1|\mathrm{H}) = .1$ and $P(\mathrm{E}_1|\neg\mathrm{H}) = .2$. Given Linda's description, it is unlikely that Linda is a bank-teller. Given that Linda's description is false, it is still unlikely that Linda is a bank-teller, but somewhat more likely. Furthermore, $P(\mathrm{E}_2|\mathrm{H}) = .6$ and $P(\mathrm{E}_1|\neg\mathrm{H}) = .1$ – given Linda's description, she is somewhat likely to be a feminist, while given that the description is false, it is less likely that she is a feminist.

Figure 5 shows that, regardless of how probable Linda's description is, using these parameters, the description (indeed) does not explain the options that need to be ranked particularly well. However, it nevertheless better explains the conjunction of Linda being a feminist and a bank-teller than Linda as a bank-teller. This makes sense: the story is strange, but it makes more sense if Linda is known to be a feminist in addition to surprisingly being a bank teller. We plan to more generally determine the conditions under which the conjunction has greater explanatory power than a single conjunct in future research.

The A-B paradigm (a): Consider the following example of A-B (from Tversky & Kahneman, 1983, p. 306):

> Peter is a junior in college who is training to run the mile in a regional meet. In his best race, earlier this season, Peter ran the mile in 4:06 min.
>
> Rank the following statements by their probability:
>
> - Peter will run the second half-mile under 1:55 min.
> - Peter will complete the mile under 4 min.
> - Peter will run the second half-mile under 1:55 min and complete the mile under 4 min.

We assume a collider network $H_1 \rightarrow E \leftarrow H_2$, where the variable H_1 stands for Peter so far always having had run a mile above 4:06 min (the description), H_2 for Peter running the second half-mile under 1:55 min (the first conjunct), and E for Peter completing the full mile in under 4 min (the second conjunct).

Our reconstruction may sound surprising at first because the description is in the role of a hypothesis. However, this is because we consider the description H_1 and the time in the second half mile H_2 as hypothetical causes of the time needed

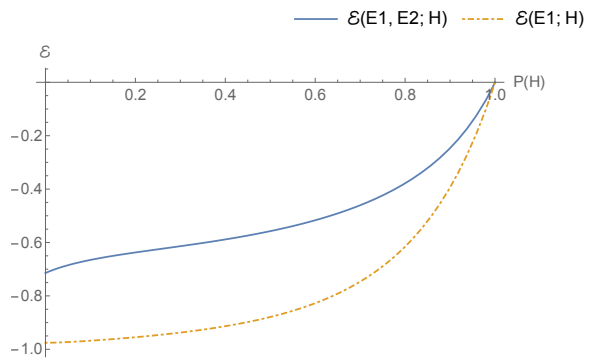

Figure 5: We assume a common cause network $E_1 \leftarrow H \rightarrow E_2$, where H stands for Linda's description, and E_1 and E_2 respectively for Linda being a bank-teller and Linda being a feminist. The parameters are set as $P(\mathrm{E_1}|\mathrm{H}) = .1$, $P(\mathrm{E_1}|\neg\mathrm{H}) = .2$, $P(\mathrm{E_2}|\mathrm{H}) = .6$ and $P(\mathrm{E_1}|\neg\mathrm{H}) = .1$. Linda's description provides a more powerful explanation of the conjunction of her being a feminist and a bank-teller (solid blue line) than of Linda as a bank-teller (dashed orange line).

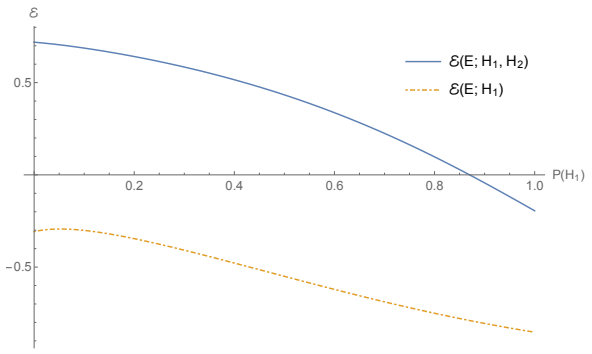

Figure 6: We assume a collider network $H_1 \rightarrow E \leftarrow H_2$, where H_1 stands for Peter so far always having had run a mile above 4:06 min (the description), H_2 for Peter running the second half-mile under 1:55 min (the first conjunct), and E for Peter completing the mile in under 4 min (the second conjunct). The parameters are set as $P(\mathrm{H_2}) = .1$, $P(\mathrm{E}|\mathrm{H_1},\mathrm{H_2}) = .6$, $P(\mathrm{E}|\mathrm{H_1},\neg\mathrm{H_2}) = .1$, $P(\mathrm{E}|\neg\mathrm{H_1},\mathrm{H_2}) = .7$, and $P(\mathrm{E}|\neg\mathrm{H_1},\neg\mathrm{H_2}) = .2$. Regardless of how likely Peter's description is, then the conjunction is more powerfully explained (solid blue line) in comparison to an explanation by the description (dashed orange line).

for the full mile, E. We proceed with some plausible parameters: $P(\mathrm{H_2}) = .1$ (it is unlikely that Peter runs the second-half mile so quickly), $P(\mathrm{E}|\mathrm{H_1},\mathrm{H_2}) = .6$ (given Peter's previous run times and a very quick second half-mile, the full mile under 4:00 minutes is somewhat likely), $P(\mathrm{E}|\mathrm{H_1},\neg\mathrm{H_2}) = .1$ (given his previous times and it not being the case that he ran the second half-mile below 1:55 min, i.e. he needed more than that, makes the full mile below 4:00 min unlikely), $P(\mathrm{E}|\neg\mathrm{H_1},\mathrm{H_2}) = .7$ (given that he had previously run the mile in under 4:06 min and that he had run the second half-mile in less than 1:55 min makes the full mile in under 4:00 min even likelier), and $P(\mathrm{E}|\neg\mathrm{H_1},\neg\mathrm{H_2}) = .2$ (given that he had already run in under 4:06 min, but that he needed more than 1:55 min for the second half-mile, a full mile under 4:00 min turns out to be unlikely).

We can then show that regardless of how likely Peter's description is, the conjunction turns out to always be preferable in comparison to an explanation of running under 4:00 min by the description on its own, hence we can rightfully expect the CF. See Figure 6. As in the M-A cases, we plan to determine more general properties that tip the model towards or against the CF in the A-B scenarios too.

The A-B paradigm (b): In case the description is irrelevant (e.g., in the case of Mr. F who is randomly selected), it is clear that the evidence (stroke) is better explained by two hypotheses: that Mr. F. was selected at random and that he is over 55 years old than by Mr. F. being selected at random alone. We omit the details as this follows straight-forwardly. Hence, our explanation-based account masters both main types of the CF. And although there are many variations of the CF which we did not address due to space limitations and even though our explanation-based account of the CF may turn out to not be able to adequately cover all possible scenarios, our contribution provides a promising starting point for further theoretical and experimental research on how explanatory considerations may be involved in the CF. We can also easily explore under which conditions the fallacy is more likely to occur. Moreover, we can also consider more complex cases where double and multiple conjunction fallacies are expected. This is especially promising because if the CF is a serious challenge of human reasoning with potentially harmful practical implications, then we need to be able to explain it in complex scenarios too. An explanation of the CF via explanation as such might just do the trick.

Conclusion

In this contribution we discussed how the conjunction fallacy may be interpreted as a case of explanatory reasoning. The account is promising because it provides a probabilistic account which can explain the main types of the fallacy from the literature (the M-A and A-B paradigms). In addition, it also predicts when the fallacy is likely to occur in cases which have to the best of our knowledge not yet been investigated (probabilistic chains) and which may turn out to be problematic for some leading approaches to the CF.

Notably, this is yet another case where explanatory reasoning plausibly plays an important role (see for instance Lombrozo & Carey, 2006 for further empirical and theoretical evidence). We leave it to future research to identify what exactly it is in explanatory reasoning that makes this type of inference so pervasive that it potentially also shows up where it should not have – in assessments of probability.

Acknowledgments

This work was supported by the Arts and Humanities Research Council and The Deutsche Forschungsgemeinschaft [grant numbers HA 3000/20-1, HA 3000/21-1]. Thanks also to Katya Tentori and anonymous referees for their helpful comments.

References

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

Cevolani, G., & Crupi, V. (2022). Truth, probability, and evidence in judicial reasoning: The case of the conjunction fallacy. In *Judicial Decision-Making: Integrating Empirical and Theoretical Perspectives* (pp. 105–121). Springer.

Cohen, M. P. (2016). On three measures of explanatory power with axiomatic representations. *The British Journal for the Philosophy of Science*.

Crupi, V., Elia, F., Aprà, F., & Tentori, K. (2018). Double conjunction fallacies in physicians' probability judgment. *Medical Decision Making*, *38*(6), 756–760.

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, *14*(2), 182–199.

Crupi, V., & Tentori, K. (2012). A second look at the logic of explanatory power (with two novel representation theorems). *Philosophy of Science*, *79*(3), 365–385.

Dulany, D. E., & Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, *9*(1), 85–110.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, *50*(2), 123–129.

Fitelson, B. (1999). The plurality of bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, *66*(S3), S362–S378.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, *2*(1), 83–115.

Glass, D. H. (2002). Coherence, explanation, and Bayesian networks. In M. O'Neill, R. F. E. Sutcliffe, C. Ryan, M. Eaton, & N. J. L. Griffith (Eds.), *Artificial Intelligence and Cognitive Science, 13th Irish Conference, AICS 2002* (pp. 177–182). Berlin: Springer.

Hartmann, S., & Meijs, W. (2012). Walter the banker: The conjunction fallacy reconsidered. *Synthese*, *184*, 73–87.

Hartmann, S., & Trpin, B. (2023a). Conjunctive explanations: A coherentist appraisal. In D. H. Glass & J. N. Schupbach (Eds.), *Conjunctive Explanations* (pp. 111–142). New York and London: Routledge.

Hartmann, S., & Trpin, B. (2023b). Why coherence matters. Forthcoming in: *Journal of Philosophy*.

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*(4), 275–305.

Jönsson, M. L., & Shogenji, T. (2019). A unified account of the conjunction fallacy by coherence. *Synthese*, *196*, 221–237.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford Handbook of Thinking and Reasoning* (pp. 260–276). Oxford: Oxford University Press.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204.

Morier, D. M., & Borgida, E. (1984). The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychology Bulletin*, *10*(2), 243–252.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Olsson, E. J. (2002). What is the problem of coherence and truth? *Journal of Philosophy*, *99*(5), 246–272.

Olsson, E. J. (2022). *Coherentism*. Cambridge: Cambridge University Press.

Roche, W., & Sober, E. (2023). Purely probabilistic measures of explanatory power: A critique. *Philosophy of Science*, *90*(1), 129–149.

Schupbach, J. N., & Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science*, *78*(1), 105–127.

Shogenji, T. (1999). Is coherence truth conducive? *Analysis*, *59*(4), 338–345.

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

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science*, *28*(3), 467–477.

Tentori, K., & Crupi, V. (2012). How the conjunction fallacy is tied to probabilistic confirmation: Some remarks on schupbach (2009). *Synthese*, *184*, 3–12.

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: probability versus inductive confirmation. *Journal of Experimental Psychology: General*, *142*(1), 235.

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

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.