Notes for Week 12 of *Confirmation*

11/28/07

Branden Fitelson

1 Some Background on Subjective Bayesian Confirmation Theory

1.1 Probabilism: Epistemic *vs* Psychological

Subjective Bayesians assume that epistemic/actual agents ϕ/ψ have *degrees of confidence* or *degrees of belief* in propositions, and that these degrees of belief *satisfy the probability axioms*. In the case of *actual* agents ψ, this is surely a strong *idealization*, since it requires a fair amount of *logical omniscience* (as well as other qualities that actual agents don't, in general, possess). In the case of *epistemic* agents, probabilism is also a substantive claim, which is non-trivial to justify. In the "conjunction fallacy" (CF), both normative and descriptive questions are raised, and it is difficult to cleanly separate the two sorts of questions. I will suggest that confirmation may have a role in explaining (if not justifying) CF-responses of actual ψ's.

1.2 Background on Bayesian Relevance Measures of Confirmation

1.2.1 The Plethora of Bayesian Relevance Measures

As I have mentioned briefly before, there are various quantitative measures $\mathfrak{c}$ of (relevance) confirmation. Many such measures have been proposed and/or defended in the Bayesian literature (my dissertation *Studies in Bayesian Confirmation Theory* is all about this plethora of measures). All such measures (when properly understood) will have several things in common. First, they are all *relevance* measures. That is, they are all sensitive to probabilistic relevance. There are various ways to make this precise. One way is as follows:

$$(\mathcal{R}) \qquad \mathfrak{c}(H,E \mid K)\begin{cases} > 0 & \text{if } \Pr(H \mid E\,\&\,K) > \Pr(H \mid K), \\ < 0 & \text{if } \Pr(H \mid E\,\&\,K) < \Pr(H \mid K), \\ = 0 & \text{if } \Pr(H \mid E\,\&\,K) = \Pr(H \mid K). \end{cases}$$

This characterization of the relevance requirement presupposes that the measures are properly *scaled*, so as to respect the sign conventions in $(\mathcal{R})$. To simplify matters, I will present all the relevance measures I will be discussing on a *normalized* $[-1,1]$ *scale*. This can always be done, since all we care about (here) is the *comparative (ordinal) structure* of relevance measures. The precise *numbers* they assign will not be important for us. Before presenting and discussing some relevance measures, a definition is in order:

> **Definition**. Two measures $\mathfrak{c}_1(H,E \mid K)$ and $\mathfrak{c}_2(H,E \mid K)$ of the degree to which E confirms H relative to K are said to be *ordinally equivalent* ($\mathfrak{c}_1 \approx \mathfrak{c}_2$) just in case, for all H, E, K, H', E', K':
>
> $$\mathfrak{c}_1(H,E \mid K) \geq \mathfrak{c}_1(H',E' \mid K') \text{ iff } \mathfrak{c}_2(H,E \mid K) \geq \mathfrak{c}_2(H',E' \mid K').$$

As it turns out, all of the measures $\mathfrak{c}$ in the historical literature can be converted into ordinally equivalent measures $\mathfrak{c}'$, which are on a *normalized* $[-1,1]$ *scale*. I will give some examples below, and I will always assume we are working with *normalized* relevance measures, in this sense. The plethora is generated because there are many logically equivalent ways of saying "E and H are positively correlated, given K". For instance:

- $\Pr(H \mid E\,\&\,K) > \Pr(H \mid K)$.
- $\Pr(E \mid H\,\&\,K) > \Pr(E \mid {\sim}H\,\&\,K)$.
- $\Pr(H \mid E\,\&\,K) > \Pr(H \mid {\sim}E\,\&\,K)$.

Prima facie, this gives many ways of defining "degree of relevance confirmation", since there are many functions of the left and right hand sides of these three inequalities that satisfy $(\mathcal{R})$. Here is a sample of 5:

$$d(H,E \mid K) \stackrel{\text{def}}{=} \Pr(H \mid E\,\&\,K) - \Pr(H \mid K)$$

$$r(H,E \mid K) \stackrel{\text{def}}{=} \frac{\Pr(H \mid E \,\&\, K) - \Pr(H \mid K)}{\Pr(H \mid E \,\&\, K) + \Pr(H \mid K)}$$
$$\approx \frac{\Pr(H \mid E \,\&\, K)}{\Pr(H \mid K)}$$
$$l(H,E \mid K) \stackrel{\text{def}}{=} \frac{\Pr(E \mid H \,\&\, K) - \Pr(E \mid {\sim}H \,\&\, K)}{\Pr(E \mid H \,\&\, K) + \Pr(E \mid {\sim}H \,\&\, K)}$$
$$\approx \frac{\Pr(E \mid H \,\&\, K)}{\Pr(E \mid {\sim}H \,\&\, K)}$$
$$\approx \frac{\Pr(H \mid E \,\&\, K) \cdot [1 - \Pr(H \mid K)]}{[1 - \Pr(H \mid E \,\&\, K)] \cdot \Pr(H \mid K)}$$
$$s(H,E \mid K) \stackrel{\text{def}}{=} \Pr(H \mid E \,\&\, K) - \Pr(H \mid {\sim}E \,\&\, K)$$
$$= \frac{d(H,E \mid K)}{\Pr({\sim}E \mid K)}$$
$$z(H,E \mid K) = \begin{cases} \dfrac{d(H,E \mid K)}{\Pr({\sim}H \mid K)} & \text{if } \Pr(H \mid E \,\&\, K) \geq \Pr(H \mid K), \\ \dfrac{d(H,E \mid K)}{\Pr(H \mid K)} & \text{otherwise.} \end{cases}$$

The (interestingly, piece-wise!) z-measure has recently been defended by Crupi *et. al.* (just published in the most recent issue of *Philosophy of Science* — see below for more on their clever argument). The s-measure has been used by Joyce and Christensen (in their proposed resolutions of the old evidence problem). The l-measure has been defended by myself (and Turing, Good, and others). The r-measure has been defended by Peter Milne (and others, dating back to Johnson and Keynes). The d-measure was tentatively proposed by Carnap (2nd edition of LFP), and it has been used by many Bayesians since (including Eells and Sober).

1.2.2 Some "Normative Constraints" One Might Use to "Narrow the Field" of Relevance Measures

We have already seen that, from an *inductive-logical* point view, the following desideratum is natural:

$$(\mathcal{D}) \qquad \mathfrak{c}(H,E \mid K) = \begin{cases} 1 & \text{if } E \,\&\, K \vDash H, \\ -1 & \text{if } E \,\&\, K \vDash {\sim}H. \end{cases}$$

This is a *quantitative* version of the "deductive limiting-case" desideratum. But, this can also be stated as a *comparative* desideratum in the following way (I prefer comparative/ordinal requirements):

$(\mathcal{D}')$ For all E, H, E', H', K, and K': *if* $E \,\&\, K \vDash H$ but $E' \,\&\, K' \nvDash H'$, *then* $\mathfrak{c}(H,E \mid K) \geq \mathfrak{c}(H',E' \mid K')$.

Interestingly, only two of the five measures above satisfies $(\mathcal{D})/(\mathcal{D}')$: l and z. So, this is one way we might "narrow the field". But, it is unclear what the *normative force* of $(\mathcal{D})/(\mathcal{D}')$ is supposed to be, if $\mathfrak{c}$ is meant to be a measure of (say) *degree of evidential support*. Presumably, to give this *inductive-logical* desideratum such normative force, we would need to presuppose some sort of *bridge principle* (like RTE, for instance).

There are other desiderata that (intuitively) seem to *rule-out* some of our measures as inadequate (even when interpreted in *evidential* terms). Ellery Eells and I discuss various symmetry/asymmetry desiderata in our paper "Symmetries and Asymmetries in Evidential Support". In that paper, we argued that the following two symmetries should *not always* hold (*i.e.*, they each should *fail* for *some* E's, H's, and K's):

(CS) $$\mathfrak{c}(H,E \mid K) = \mathfrak{c}(H,E \mid K)$$

(ES) $$\mathfrak{c}(H,E \mid K) = -\mathfrak{c}(H,{\sim}E \mid K)$$

We presented counterexamples in which $E \,\&\, K \vDash H$ but $H \,\&\, K \nvDash E$. But, we argued that the entailments were not essential to the asymmetries. Here is what we said (which still sounds compelling to me):

> A card is randomly drawn from a standard deck (and K includes all the usual assumptions about such draws). Let E be the evidence that the card is the seven of spades, and let H be the hypothesis that the card is black. We take it to be intuitively clear that E is not only conclusive, but also strong, evidence in favor of H, whereas ${\sim}E$ (that the card drawn is not the seven of spades) is close to useless, or close to "informationless," with regard to the color of the card ... we have an intuitive counterexample to (ES), and we should have: $c(H,E \mid K) \gg |c(H,{\sim}E \mid K)| = -c(H,{\sim}E \mid K)$.

As stated, this example seems to be trading on desiderata $(\mathcal{D})/(\mathcal{D}')$. But, as we point out in the paper:

> ... the conclusiveness feature of the example (that E logically implies H) is not what is at the heart of the counterexample. To see this, simply consider a modification of the example where E is a report of suit/rank from a highly reliable, but fallible, assistant.

Moreover, this example can simultaneously be seen as a counterexample to (CS) as well. As we explain:

> It seems clear from these examples that a piece of evidence E can confirm a hypothesis H to a much different degree than H confirms E. Consider for example whether the observation that a card is the seven of spades confirms the proposition that the card is black equally well as the proposition that the card is black confirms the proposition that the card is the seven of spades. With initial uncertainty about the value of the card, we consider the seven of spades, as evidence, to be more highly informative and confirmatory of the blackness of the card, as hypothesis, than the blackness of the card, as evidence, is for the card's being the seven of spades in particular.

So, it seems that (ES) and (CS) must *not* hold, for *all* E, H, and K. This is bad news for r and s, since r entails that (CS) holds universally, and s entails that (ES) holds universally (that's a simple exercise in probability). As such, it seems that r and s are *defective*, from an *epistemic* (or *logical*) point of view. In our paper, we also suggested that some symmetries *should* hold universally. For instance, we endorsed (without argument):

(HS) $$\mathfrak{c}(H, E \mid K) = -\mathfrak{c}(\sim H, E \mid K)$$

Of our five measures, all of them *except* r entails that (HS) holds universally. Since we already have a knock-down argument against r, this positive desideratum doesn't actually do very much work for us. In general, we were more tentative about such *positive* desiderata in that paper (and I remain so). However, Crupi *et. al.* have picked-up where we left off. They suggest that, if one follows our line of argument to its logical conclusion, then one ends-up with a set of positive and negative symmetry/asymmetry desiderata that rules-out all of these measures except z (in fact, they prove a much stronger *uniqueness* result for z). I won't get into all the details of their elegant general framework for thinking about such symmetries and asymmetries. But, I will point out the key observation they make which rules-out l in favor of z (and which hadn't occurred to me as a possibility — I'm still not sure exactly what we should say about it). Crupi *et. al.* point out that, while (CS) does not make sense in the case of *confirmatory* E, our own framework of symmetries/asymmetries (in its full "lattice-theoretic glory") seems to suggest the following:

(CS′) *If* E *dis*confirms H, relative to K, *then* $\mathfrak{c}(H, E \mid K) = \mathfrak{c}(E, H \mid K)$.

In other words, the framework of our paper suggests an *asymmetry* between confirmatory and disconfirmatory examples, when it comes to (CS). The reason for this [again inspired in the background by $(\mathcal{D})/(\mathcal{D}')$] is that while $E \models H \not\Rightarrow H \models E$, it *is* the case that $E \models \sim H \Rightarrow H \models \sim E$. As a result, as soon as one imagines a case in which E is conclusive for $\sim H$, this will also be a case in which H is conclusive for $\sim E$. And, because we accept (HS), we will then be forced to see any such example as a *non*-counterexample to (CS). Thus, if we focus on the deductive limiting-cases, we get an interesting asymmetry here between disconfirmation and confirmation [remember Nicod's emphasis on related asymmetries?]. And, following the general line of argument sketched (incompletely) in our paper, (CS′) can then be seen as a natural *positive* symmetry desideratum, even though (CS) fails for some *confirmatory* cases. This is a very interesting (and clever) point. And, it explains why z is defined (piece-wise) in different ways for confirmatory *vs* disconfirmatory cases. While I find their paper fascinating and illuminating, I am not completely convinced, for a couple of reasons. First, in our paper, we didn't argue for any positive symmetry desiderata, and this was because we didn't know of any compelling arguments (I still don't). Thus, I prefer to endorse *only* the *negative* desiderata discussed in that paper (which are all *common ground* between l and z). Second, there are *other* desiderata (which I will not discuss here) that z *fails* to satisfy, *because* of its piece-wise definition. Nonetheless, I highly recommend the Crupi *et. al.* (which I have now posted, along with the original Eells & Fitelson paper). I also recommend the Tentori *et. al.* paper (*Cognition* 2007), which compares various $\mathfrak{c}$'s with respect to *descriptive* adequacy.

So, what does all of this have to do with the conjunction fallacy (CF)? Well, as it turns out, there have been several attempted confirmation-theoretic treatments of the conjunction fallacy. But, they have all used measures such as d, r, or s, which we think are *normatively defective*. As a result, if we want a "*normatively compelling* explanation" of the response patterns in conjunction fallacy experiments, then we'd better come up with another approach. Indeed, the approach I will now describe (forthcoming in *Thinking and Reasoning* — Crupi, Fitelson, and Tentori) is *completely robust* — it goes through for *any* of the relevance measures of confirmation we've been discussing (indeed, *for all relevance measures I have ever seen*).

2 The Conjunction Fallacy

2.1 The Original Example and Some Possible "Exculpatory Moves"

Tversky and Kahneman (1983) discuss the following example, which was the first example of the "conjunction fallacy". In T&K's experiment, subjects are given the following *evidence* about person named Linda:

(*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.

Then, the subjects are asked (some version of) the following question:

(Q) Is it more probable (assuming E is true) that Linda is ("H_1") a bank teller, or ("H_1 and H_2") a bank teller *and* an active feminist?

Most subjects (ψ) say that "H_1 and H_2" is more probable (given E) than "H_1". On its face, this violates comparative probability theory, since $X \models Y$ implies $\Pr(X \mid E) \leq \Pr(Y \mid E)$, and $H_1 \,\&\, H_2 \models H_1$. There are some obvious moves one might try here in order to "rationalize" such responses. Here are two popular moves:

1. Perhaps there is some *pragmatic interference* here, and subjects actually understand the contrast as *not* one between H_1 and $H_1 \,\&\, H_2$. Maybe they contrast $H_1 \,\&\, {\sim}H_2$ with $H_1 \,\&\, H_2$. Or, maybe they don't understand H_1 to be a *conjunct* of "H_1 and H_2", *etc.* If the contrast is understood in such a way that the two elements of the contrast are not logically related, then no "fallacy" need be committed at all.

    - There have been many additional experiments performed which control for pragmatic factors, and ensure that the subjects understand that the contrast is between H_1 and $H_1 \,\&\, H_2$. And, the patterns of response are almost identical. See website for several papers on this.

2. Maybe the subjects aren't understanding the question as a question about *conditional probabilities*.

    - Many experiments have been performed to control for this too. My favorite ones use *betting odds* (in various ways). Also, the question can be placed in various sorts of alternative conditional forms, *etc.*, and in ways that make the supposition of E salient in different ways, *etc.* Same patterns. Again, see the website for papers which discuss such alternative experiments.

Because of these many experiments that have been done since T&K's original paper, I am inclined to agree that people are not giving "correct" responses to the question that is asked. T&K and their students have developed *non*-probabilistic theories of "confidence", in order to try to explain what is going on here. I (we) think confirmation theory can be useful here — not to try to argue that the responses are "correct", but to show that there might be *probabilistic* (but *confirmation-theoretic*) features of the example that mirror the structure of the response patterns observed. In other words, the aim is to use confirmation theory to identify a ("rational") feature of the example that *does* have the same structure as the most common response.

2.2 Confirmation Theory and the Conjunction Fallacy

Isaac Levi (in his review of T&K's paper) was probably the first to suggest that the confirmation-theoretic structure of the Linda example may shed some light on what's fooling the subjects. Since then, several authors have looked at using relevance measures of confirmation to examine the confirmation-theoretic structure of the Linda example (several exemplars, including a more recent paper by Levi, are posted on the website). However, these analyses have all been piece-meal, and they have relied on measures of confirmation that we think are *normatively defective*. We give a completely *robust* (*measure-insensitive*) analysis.

Before explaining how our analysis works, let's remind ourselves of some history. Recall that Carnap showed (1950) that, while (SCC) *holds* for *firmness* confirmation (confirmation$_f$), it *fails* for *increase-in-firmness* (*relevance*) confirmation (confirmation$_i$). That is, we have the following two facts:

(SCC$_f$) If E confirms$_f$ H, then E *must* also confirm$_f$ any logical consequence H' of H.

${\sim}$(SCC$_i$) If E confirms$_i$ H, then E *need not* confirm$_i$ all logical consequences H' of H.

– Simple counterexample for confirms$_i$: sampling a card at random from a standard deck. $E \stackrel{\text{def}}{=}$ the card is black, $H \stackrel{\text{def}}{=}$ the card is the ace of spades, $H' \stackrel{\text{def}}{=}$ the card is an ace.

So, it has long been known that such things are *possible*, from a *qualitative* point of view. But, the question at hand involves a *comparative* claim, *not* a *qualitative* one. What we seek is a filling-in of the "Linda story", which undergirds the following comparative claim — ideally, for *any* relevance measure $\mathfrak{c}$:

(†) $\mathfrak{c}(H_1 \,\&\, H_2, E) > \mathfrak{c}(H_1, E)$. [Here, we suppress the background corpus K, for simplicity.]

This (†) is, of course, impossible for any firmness measure of confirmation (which will just be a conditional probability function). But, this *is* possible for *relevance* measures. The question is: can we identify a set of conditions on $\mathfrak{c}$ that are (intuitively) satisfied in the Linda case, and which will entail (†) for all (or a sufficiently broad class of) relevance measures? The answer is: YES! Here is one such a set of conditions:

Theorem 1. For "*all*" Bayesian relevance measures $\mathfrak{c}$, *if*

(i) $\mathfrak{c}(H_2, E \mid H_1) > 0$ *and*

(ii) $\mathfrak{c}(H_1, E) \leq 0$,

then $\mathfrak{c}(H_1 \,\&\, H_2, E) > \mathfrak{c}(H_1, E)$.

By "*all*", I mean all relevance measures satisfying what Joyce calls the *Weak Law of Likelihood* (in his *Bayes's Theorem* Entry in the *Stanford Encyclopedia*), which is the following constraint:

(WLL) If $\Pr(E \mid H_1) > \Pr(E \mid H_2)$ and $\Pr(E \mid {\sim}H_1) \leq \Pr(E \mid {\sim}H_2)$, then $\mathfrak{c}(H_1, E) > \mathfrak{c}(H_2, E)$.

All relevance measures that have appeared in the historical literature satisfy (WLL). Indeed, I think of (WLL) as a *universal desideratum for any relevance measure* $\mathfrak{c}$ — that is, I think of (WLL) as being on a par with ($\mathcal{R}$) itself. Here is an alternative set of conditions which also entails (†):

Theorem 2. For "*all*" Bayesian relevance measures $\mathfrak{c}$, *if*

(i) $\Pr(E \mid H_1 \,\&\, {\sim}H_2) < \Pr(E \mid H_1 \,\&\, H_2)$ *and*

(ii′) $\Pr(E \mid H_1 \,\&\, {\sim}H_2) \leq \Pr(E \mid {\sim}H_1 \,\&\, H_2)$,

(ii″) $\Pr(E \mid H_1 \,\&\, {\sim}H_2) \leq \Pr(E \mid {\sim}H_1 \,\&\, {\sim}H_2)$,

then $\mathfrak{c}(H_1 \,\&\, H_2, E) > \mathfrak{c}(H_1, E)$.

Note that the first condition (i) here is *logically equivalent to* the first condition (i) in Theorem 1. The second, probabilistic formulation might be more intuitive to verify in the Linda case. Our subjects (and everyone I've ever asked!) universally assent to the truth of *all* of these four conditions — stated *either* in "confirmation" ("evidential support") language, *or* in conditional probability language. Indeed, it seems obvious to us that all of these conditions should be assented to in the Linda case. After all, it seems to us that they are all *intuitively correct* (given typical background knowledge of a typical subject, which we've suppressed here). That is, we think it would be rational for a typical agent in this experiment to accept all of these conditions.

We are *not* claiming that the subjects are "really giving the correct answer after all". On this point, we agree with T&K (especially in light of all the subsequent experiments). However, we think it is interesting that there is a ("epistemically rational") *probabilistic feature* of the example that has the same comparative structure as the most common response to (Q). Recall that Carnap himself was sometimes confused about the distinction between firmness and increase-in-firmness. So, why couldn't our subjects be similarly confused? We think this is a neat explanation of the responses. It doesn't just give a "merely descriptive" model of the agents. Rather, it is an explanation that points to a *prescriptively interesting*, *probabilistic* feature (confirmation$_i$) of the example that is easily confused with the (confirmation$_f$) features involved in (Q). Interestingly, T&K themselves said something which is suggestive of this explanation (Levi picked-up on it):

> . . . feminist bank teller is a better hypothesis about Linda than bank teller.

We agree! And, by this, *we* mean that *E confirms$_i$ H_1&H_2 more strongly than* H_1, as ensured by our Theorems 1 and 2. In the next section, I'll briefly discuss a different sort of example which also seems to exhibit a "CF-effect", but which does not fit the mold of our Theorems 1 and 2. This is work in progress. We're trying to think-up (a different set of) conditions which would entail (†) for these cases. We welcome suggestions!

2.3 A Different Class of "Conjunction Fallacy" Cases

- Inequalities equivalent to our (i) [$\Pr(E \mid H_1 \,\&\, H_2) > \Pr(E \mid H_1)$] have been empirically vindicated in traditional CF cases [14].
- Our (ii)'s have not been explicitly tested. But, we suspect these will obtain (empirically) in the the traditional CF cases.
- We are performing experiments to test these (i)/(ii) and (i)/(ii′)/(ii′′) accounts of the traditional CF cases [4].
- Interestingly, there are other (non-traditional) sorts of CF cases in which the (ii)'s seem much less intuitive. *E.g.* [16]

  E = John is Scandanavian.[4] H_1 = John has blue eyes.
  H_2 = John has blond hair.
- This example also seems to exhibit a conjunction fallacy pattern (this time, symmetric with respect to H_1 and H_2).
- 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.

[4] Scandinavia has the greatest percentage of people with blond hair and blue eyes (though every possible combination of hair and eye color occurs there). E = John is sampled at random from the Scandinavian population.

- In the John case, (i) seems intuitively plausible, but (ii) does not. What about (ii′) and (ii′′)? While (ii′) is not patently false (the = case, anyway), (ii′′) seems pretty clearly false.
- Moreover, (i) *alone* cannot undergird a *robust* account. It only suffices for *some* $\mathfrak{c}$'s, which lack normative force [6].
- *Descriptively*, we suspect confirmation-theoretic relations between H_1 and H_2 *themselves* may be involved in this CF.
- Specifically, the terms $\mathfrak{c}(H_i, H_j)$ seem to be salient. We bet they are *explanatorily* relevant. There is some preliminary evidence in the literature for this (but no general model).
- Psychologically, we think there are two important sets of confirmation-theoretic factors involved in CF cases:
  - $\mathfrak{c}(H_1, E)$, $\mathfrak{c}(H_2, E)$, $\mathfrak{c}(H_1, E \mid H_2)$, $\mathfrak{c}(H_2, E \mid H_1)$. [Traditional CF]
  - $\mathfrak{c}(H_1, H_2)$, $\mathfrak{c}(H_2, H_1)$, $\mathfrak{c}(H_1, H_2 \mid E)$, $\mathfrak{c}(H_2, H_1 \mid E)$. [NT CF]
- We are working on more general confirmaiton-theoretic models which we hope will account for (*i.e.*, predict and explain, if not rationalize) all known conjunction fallacies.