The Reasoner, 6 (3), 2012: 39-40
[Erratum in *The Reasoner*, 6 (4), 2012: 68]

Vincenzo Crupi

An argument for not equating confirmation and explanatory power

Quantitative explicata of confirmation and of explanatory power are often defined as functions of probability values involving an event *e* and a hypothesis *h*. These two kinds of formal constructs also tend to display common features and salient analogies (see Schupbach J. and Sprenger J., 2011, "The logic of explanatory power", *Philosophy of Science*, 78, pp. 105-27, and Crupi V. and Tentori K., 2012, "A second look at the logic of explanatory power", *Philosophy of Science*, 79, pp. 365-385, for some relevant remarks and further references). In view of this, investigating the connection between a probabilistic measure of confirmation $C(h,e)$ and of explanatory power $E(e,h)$ appears appropriate, if not pressing, as a source of theoretical clarification. An instructive possibility to explore is the statement of outright identity between the two notions. As a concrete illustration, consider that I.J. Good's formal representation of the explanatory power of candidate explanans *h* with regards to explanandum *e* (in "Weight of evidence, corroboration, explanatory power, information and the utility of experiments", *Journal of the Royal Statistical Society, Series B*, 1960, 22, pp. 319-31) is identical to the "one true measure" of the degree of confirmation from *e* to *h* as once advocated by Peter Milne ("Log[$P(h/eb)/P(h/b)$] is the one true measure of confirmation", *Philosophy of Science*, 1996, 63, pp. 21-6) – of course with the caveat that the hypothesis *h* at issue be in some explanatory relation with evidence *e* at all. Always keeping this latter proviso in mind, the general form of a "reductionist" claim to identity would be as follows:

Reduction (*R*).
For any e,h and any P, $C(h,e) = E(e,h)$.

(I'll be assuming throughout that statements are contingent and the probability function P is regular.) Statement *R* may seem overly strong to begin with, but notice that it neatly conveys so-called *inference to the best explanation* (IBE). After all, for advocates of the IBE view, "observations support the hypothesis *precisely because* it would explain them" (Lipton P., "Inference to the best explanation", in W.H. Netwon-Smith, ed., *A Companion to the Philosophy of Science*, Blackwell, 2000, p. 185, emphasis added). And *R* is of concern even beyond that, if only because it would arguably trivialize the division of labour between two otherwise distinct branches of formal epistemology and philosophy of science. For short, *R* is not to be dismissed too quickly, i.e., unless a relevant argument is provided to undermine it. Such an argument is put forward in what follows.

A compelling principle of a model of explanatory power seems to be that the better hypothesis *h* would succeed in explaining the occurrence of a state of affairs *e* the worse it would fail in explaining the occurrence of its complementary, $\neg e$. Formally, such an inverse ordinal correlation between explanatory success and explanatory failure with regards to a pair of complementary statements *e* and $\neg e$ is spelled out as follows:

Symmetry (*S*).
For any e_1,e_2,h and any P, $E(e_1,h) \gtreqless E(e_2,h)$ iff $E(\neg e_1,h) \lesseqgtr E(\neg e_2,h)$.

On the other hand, consider the following condition concerning confirmation:

Final probability incrementality (*F*).
For any h,e_1,e_2 and any P, $C(h,e_1) \gtreqless C(h,e_2)$ iff $P(h|e_1) \gtreqless P(h|e_2)$.

Condition *F* states that, for any given hypothesis h, confirmation is an increasing function of the posterior probability conditional on the evidence at issue – a virtually unchallenged assumption in contemporary probabilistic analyses of confirmation.
Notably, the following can be proven (see below):

Theorem. {*S*,*F*} is consistent, but {*R*,*S*,*F*} is not.

Relying on both *S* and *F* as sound, the theorem above discredits the reductionist claim to identity *R*. Apparently, probabilistic confirmation and explanatory power cannot be identified, for the two notions are constrained by genuinely distinct principles on a quite basic level. They are irreducible, or "independent", much in the sense of Peano's disciple Padoa (on so-called Padoa's method in axiomatics, see Suppes P., *Introduction to Logic*, Van Nostrand, 1957, pp. 169 ff.). For all its tempting simplicity, thus, the reductionist thesis *R* turns out to be a naïve view of the connection between confirmation and explanatory power. This is not to say, of course, that there cannot be other meaningful and systematic relationships. This does mean, however, that one natural candidate formal rendition of IBE is flawed.

Proof of the Theorem.

For the first clause of the theorem, posit $C(h,e) = P(h|e) - P(h)$ and $E(e,h) = P(e|h) - P(e)$. This demonstrably makes *S* and *F* jointly true, so {*S*,*F*} is consistent. [Note: This is only one choice of measures yielding the relevant result, namely, that *S* and *F* be jointly satisfied. Another nice example is as follows: $C(h,e) = P(h|e)/P(h|\neg e)$ and $E(e,h) = P(e|h)/P(e|\neg h)$.]

For the second clause of the theorem, let P be such that $P(e|h) \neq P(e)$ and x is probabilistically independent from e, h, and their conjunction. We thus have $P(h\wedge e)/P(e) = P(h\wedge e)P(x)/P(e)P(x) = P(h\wedge e\wedge x)/P(e\wedge x)$, so that:

(*) $P(h|e) = P(h|e\wedge x)$

We also have $[P(e|h) - P(e)]P(x) \neq P(e|h) - P(e)$, which holds iff $P(e|h)P(x) - P(e)P(x) \neq P(e|h) - P(e)$ iff $-P(e|h) - P(e)P(x) \neq -P(e) - P(e|h)P(x)$ iff $1 - P(e|h) - P(e)P(x) + P(e)P(x)P(e|h) \neq 1 - P(e) - P(e|h)P(x) + P(e)P(x)P(e|h)$ iff $[1 - P(e|h)][1 - P(e)P(x)] \neq [1 - P(e)][1 - P(e|h)P(x)]$ iff $[1 - P(e|h)]/[1 - P(e)] \neq [1 - P(e|h)P(x)]/[1 - P(e)P(x)]$ iff $[1 - P(e|h)]/[1 - P(e)] \neq [1 - P(e\wedge x|h)]/[1 - P(e\wedge x)]$ iff $P(\neg e|h)/P(\neg e) \neq P(\neg(e\wedge x)|h)/P(\neg(e\wedge x))$ iff $P(\neg e|h)P(h)/P(\neg e) \neq P(\neg(e\wedge x)|h)P(h)/P(\neg(e\wedge x))$, that is (by Bayes's theorem):

(**) $P(h|\neg e) \neq P(h|\neg(e\wedge x))$

Now, by *F* and (*), $C(h,e) = C(h,e\wedge x)$. Thus, by *R*, $E(e,h) = E(e\wedge x,h)$. By *S*, the latter implies $E(\neg e,h) = E(\neg(e\wedge x),h)$, so that, again by *R*, $C(h,\neg e) = C(h,\neg(e\wedge x))$. Yet *F* and (**) imply the opposite, i.e., $C(h,\neg e) \neq C(h,\neg(e\wedge x))$. So {*R*,*S*,*F*} is inconsistent. QED.