Clear Thinking in an Uncertain World: Human Reasoning and its Foundations

Lecture 6

Eric Pacuit

Department of Philosophy
University of Maryland, College Park
pacuit.org
epacuit@umd.edu

October 7, 2013

Wason Selection Task and the Paradox of Confirmation

L. Humberstone. *Hempel Meets Wason*. Erkenntnis 41 (1994), 391 - 402.

B. Fitelson and J. Hawthorne. *The Wason Selection Task(s) and the Paradox of Confirmation*. Philosophical Perspectives, Volume 24, Issue 1, pages 207 - 241, 2010.

Background Remarks, I

- E **entails** H,

Background Remarks, I

- E **entails** H,
- E **confirms** H,

Background Remarks, I

- E **entails** H,
- E **confirms** H,
- E provides **evidential support** for H

Background Remarks, II

Carnap's desiderata for inductive logic/confirmation theory:

- Confirmation theory aims to characterize a function $\mathfrak{c}(H, E)$, which generalizes entailment, in the sense that $\mathfrak{c}(H, E)$ should take on a maximal value when $E \models H$ and a minimal value when $E \models \neg H$.

Background Remarks, II

Carnap's desiderata for inductive logic/confirmation theory:

- Confirmation theory aims to characterize a function $\mathfrak{c}(H, E)$, which generalizes entailment, in the sense that $\mathfrak{c}(H, E)$ should take on a maximal value when $E \models H$ and a minimal value when $E \models \neg H$.
- The relation $\mathfrak{c}$ should be objective and logical.

Background Remarks, II

Carnap's desiderata for inductive logic/confirmation theory:

- Confirmation theory aims to characterize a function $\mathfrak{c}(H, E)$, which generalizes entailment, in the sense that $\mathfrak{c}(H, E)$ should take on a maximal value when $E \models H$ and a minimal value when $E \models \neg H$.
- The relation $\mathfrak{c}$ should be objective and logical.
- Confirmation theory/inductive logic should be applicable to/connected with epistemology in some (non-trivial) way.

Background Remarks, II

Carnap's desiderata for inductive logic/confirmation theory:

- Confirmation theory aims to characterize a function $\mathfrak{c}(H, E)$, which generalizes entailment, in the sense that $\mathfrak{c}(H, E)$ should take on a maximal value when $E \models H$ and a minimal value when $E \models \neg H$.
- The relation $\mathfrak{c}$ should be objective and logical.
- Confirmation theory/inductive logic should be applicable to/connected with epistemology in some (non-trivial) way.
- The relation $\mathfrak{c}$ should be defined in terms of probability

Background Remarks, II

Carnap's desiderata for inductive logic/confirmation theory:

- Confirmation theory aims to characterize a function $\mathfrak{c}(H, E)$, which generalizes entailment, in the sense that $\mathfrak{c}(H, E)$ should take on a maximal value when $E \models H$ and a minimal value when $E \models \neg H$.
- The relation $\mathfrak{c}$ should be objective and logical.
- Confirmation theory/inductive logic should be applicable to/connected with epistemology in some (non-trivial) way.
- The relation $\mathfrak{c}$ should be defined in terms of probability

Background Remarks, III

(LP) Everything follows from an inconsistent set of statements.

Background Remarks, III

(LP) Everything follows from an inconsistent set of statements.

(BP) If an agent's beliefs are inconsistent, then the agent (should) believe everything.

Background Remarks, IV

Interpretations of Probability:

1. A quasi-logical concept, which is meant to measure objective evidential support relations. For example, in light of the relevant seismological and geological data, it is probable that California will experience a major earthquake this decade.
2. The concept of an agent's degree of confidence, a graded belief. For example, I am not sure that it will rain in Canberra this week, but it probably will.
3. An objective concept that applies to various systems in the world, independently of what anyone thinks. For example, a particular radium atom will probably decay within 10,000 years.

A. Hájek. *Interpretations of Probability*. The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (ed.).

The Ravens Paradox

(NC) For all names a and for all (classically) logically independent predicate expressions φ and ψ,
$\varphi(a) \wedge \psi(a)$ confirms $\forall x(\varphi(x) \rightarrow \psi(x))$

(EQC) For all statements, E, H, and H', if E confirms H
and H is logically equivalent to H', then E also confirms H'.

The Ravens Paradox

(NC) For all names a and for all (classically) logically independent predicate expressions φ and ψ,
$\varphi(a) \wedge \psi(a)$ confirms $\forall x(\varphi(x) \rightarrow \psi(x))$

(EQC) For all statements, E, H, and H', if E confirms H and H is logically equivalent to H', then E also confirms H'.

$B(x)$:= 'x is black' $R(x)$:= 'x is a Raven'

(1) $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(\neg B(x) \rightarrow \neg R(x))$

(2) $\forall x(\neg B(x) \rightarrow \neg R(x))$ is logically equivalent to $\forall x(R(x) \rightarrow B(x))$

(3) $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(R(x) \rightarrow B(x))$

(1) $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(\neg B(x) \rightarrow \neg R(x))$

(2) $\forall x(\neg B(x) \rightarrow \neg R(x))$ is logically equivalent to $\forall x(R(x) \rightarrow B(x))$

(3) $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(R(x) \rightarrow B(x))$

But, then does a white jacket confirm all Ravens are black?

(1) $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(\neg B(x) \rightarrow \neg R(x))$

(2) $\forall x(\neg B(x) \rightarrow \neg R(x))$ is logically equivalent to $\forall x(R(x) \rightarrow B(x))$

(3) $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(R(x) \rightarrow B(x))$

But, then does a white jacket confirm all Ravens are black? *Not really*:

$W(a) \wedge J(a)$ confirms $\forall x(R(x) \rightarrow B(x))$ does not *follow* from (3)

(1) $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(\neg B(x) \rightarrow \neg R(x))$

(2) $\forall x(\neg B(x) \rightarrow \neg R(x))$ is logically equivalent to $\forall x(R(x) \rightarrow B(x))$

(3) $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(R(x) \rightarrow B(x))$

But, then does a white jacket confirm all Ravens are black? *Not really*:

$W(a) \wedge J(a)$ confirms $\forall x(R(x) \rightarrow B(x))$ does not *follow* from (3)

We need: $W(a)$ entails $\neg B(a)$ and $J(a)$ entails $\neg R(a)$ plus

(M) For all names a, for all (classically) consistent predicates φ and ψ, and for all statements H: If $\varphi(a)$ confirms H, then $\varphi(a) \wedge \psi(a)$ confirms H.

What is the *bridge principle* connecting the *logical* relationship E confirms H and the *epistemological* relationship E provides evidential support for H?

What is the *bridge principle* connecting the *logical* relationship E confirms H and the *epistemological* relationship E provides evidential support for H?

Consider the analogy with the fact that anything follows from an inconsistent set of sentences (called "explosion"): "One might think that it would sanction arbitrary *inferences* from inconsistent sets of *beliefs*. But this requires a *bridge principle* to connect logic and epistemology.

(recall Harman's analysis pointing out that logic alone doesn't tell us which inferences are kosher and which are not)

...the fact that Hempel's logical relation of confirmation has the property that $\neg B(a) \wedge \neg R(a)$ *(or even* $W(a) \wedge J(a)$*) confirms* $\forall x(R(x) \rightarrow B(x))$ *is only problematic to the extent that we conflate this logical claim with some epistemic claim like "observing white jackets is a way of obtaining evidence that is relevant to the claim that all ravens are black".* *(pg. 3)*

Explaining away the paradox

Hempel suggests that confirmation should be thought of as a three-place relation: E confirms H, relative to a background corpus K.

(3) $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(R(x) \rightarrow B(x))$ relative to $\top$

(3^*) $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(R(x) \rightarrow B(x))$ relative to $\neg R(a)$

Hempel suggests that people who find (3) unintuitive are conflating (3) with (3^*), and this is why they are (mis)lead to suspect that (3) is false.

($\mathcal{E}$) If S already knows that a is a non-raven, then S's observing a's color will not generate any evidence (for S) about the color of ravens. But, if S knows nothing about a, then S's learning (say, by observation of a) that $\neg B(a) \wedge \neg R(a)$ does provide some evidence (for S) that all ravens are black.

Bridge Principle

(BP) E evidentially supports H for S (in a context C) iff E confirms H, relative to K, where K is S's *total evidence* in context C.

Bridge Principle

(BP) E evidentially supports H for S (in a context C) iff E confirms H, relative to K, where K is S's *total evidence* in context C.

($\mathcal{E}'$) If $K \models \neg R(a)$, then $\neg B(a) \wedge \neg R(a)$ *does not* confirm $\forall x(R(x) \rightarrow B(x))$, relative to K. But if $K = \top$, then $\neg B(a) \wedge \neg R(a)$ confirms $\forall x(R(x) \rightarrow B(x))$, relative to K.

Monotonicity, again

(Mon) E confirms H, relative to $\top$ implies E confirms H relative to any K (provided that K does not mention any individuals not already mentioned in E).

Probabilistic Approach to the Paradox of Confirmation

(Qualitative Confirmation) E confirms H, relative to K iff
$Pr(H \mid E \ \& \ K) > Pr(H \mid K)$
where $Pr(\cdot \mid \cdot)$ is some suitable conditional probability function.

Probabilistic Approach to the Paradox of Confirmation

(Qualitative Confirmation) E confirms H, relative to K iff
$Pr(H \mid E \ \& \ K) > Pr(H \mid K)$
where $Pr(\cdot \mid \cdot)$ is some suitable conditional probability function.

$\mathfrak{c}(H, E \mid K)$: "The degree to which E is probabilistically (confirmationally) relevant, conditional on K"

Probabilistic Approach to the Paradox of Confirmation

(Qualitative Confirmation) E confirms H, relative to K iff
$Pr(H \mid E \ \& \ K) > Pr(H \mid K)$
where $Pr(\cdot \mid \cdot)$ is some suitable conditional probability function.

$\mathfrak{c}(H, E \mid K)$: "The degree to which E is probabilistically (confirmationally) relevant, conditional on K"

(Comparative) E_1 confirms H (relative to K) more strongly than E_2 confirms H (relative to K) iff
$Pr(H \mid E_1 \ \& \ K) > Pr(H \mid E_2 \ \& \ K)$.

Denying NC

(NC) For all names a and for all (classically) logically independent predicate expressions φ and ψ, $\varphi(a) \wedge \psi(a)$ confirms $\forall x(\varphi(x) \rightarrow \psi(x))$

Denying NC

(NC) For all names a and for all (classically) logically independent predicate expressions φ and ψ, $\varphi(a) \wedge \psi(a)$ confirms $\forall x(\varphi(x) \rightarrow \psi(x))$

Let K be: Exactly one of the following two hypotheses is true: (H) there are 100 black ravens, no nonblack ravens, and 1 million other things in the universe or ($\neg H$) there are $1,000$ black ravens, 1 white raven, and 1 million other things in the universe.

Denying NC

(NC) For all names a and for all (classically) logically independent predicate expressions φ and ψ, $\varphi(a) \wedge \psi(a)$ confirms $\forall x(\varphi(x) \rightarrow \psi(x))$

Let K be: Exactly one of the following two hypotheses is true: (H) there are 100 black ravens, no nonblack ravens, and 1 million other things in the universe or $(\neg H)$ there are $1,000$ black ravens, 1 white raven, and 1 million other things in the universe. Let E be $R(a)$ & $B(a)$ (with a randomly sampled from the universe). Then:

$$Pr(E \mid H \ \& \ K) = \frac{100}{1000100} \ll \frac{1000}{1001001} = Pr(E \mid \neg H \ \& \ K)$$

Denying NC

(NC) For all names a and for all (classically) logically independent predicate expressions φ and ψ, $\varphi(a) \wedge \psi(a)$ confirms $\forall x(\varphi(x) \rightarrow \psi(x))$

Let K be: Exactly one of the following two hypotheses is true: (H) there are 100 black ravens, no nonblack ravens, and 1 million other things in the universe or $(\neg H)$ there are $1,000$ black ravens, 1 white raven, and 1 million other things in the universe. Let E be $R(a)$ & $B(a)$ (with a randomly sampled from the universe). Then:

$$Pr(E \mid H \ \& \ K) = \frac{100}{1000100} \ll \frac{1000}{1001001} = Pr(E \mid \neg H \ \& \ K)$$

Therefore, E lowers the probability of (viz. *disconfirms*) H, relative to K.

Hempel: (NC) and (3) were only meant to be asserted *relative to tautological or empty background corpus*.

Hempel: (NC) and (3) were only meant to be asserted *relative to tautological or empty background corpus*.

But: what does it *mean* to talk about "the probability of H relative to *tautological* or *empty* background corpus? This requires comparing $Pr(H \mid E)$ with $Pr(H \mid \top)$ for some "suitable" conditional probability measure $Pr(\cdot \mid \cdot)$.

Maher's counterexample to NC

According to standard logic, 'All unicorns are white' is true if there are no unicorns. Given what we know, it is almost certain that there are no unicorns and hence 'All unicorns are white; is almost certainly true. But now imagine that we discover a white unicorn; this astounding discovery would make it no longer so incredible that a non-white unicorn exists and hence would disconfirm [lower the probability of] 'All unicorns are white.'

Logical probabilities vs. subjective probabilities

"That is, $Pr(H \mid E)$ now gets interpreted as (something like) the degree of belief (or degree of confidence) that S assigns to H, on the supposition that E is true....In either case, these "subjective" probabilities are much more psychologistic than the Carnapian (or Hempelian) confirmation relations weve been talking about so far."

Descriptive vs. Prescriptive Analyses

Philosophical discussions of the Paradox of Confirmation are, presumably, not offering mere descriptions of attitudes actual people happen to have about ravens, etc. Rather, they are trying to argue that various attitudes people have to the Paradox either are (or are not) reasonable or rational.

Contemporary Bayesian approaches to The Paradox are more subtle in their aims. They are not trying to argue for (or rationalize) the acceptability or unacceptability of the qualitative confirmation-theoretic claim (PC). As we explained above, that (Hempelian) debate is considered otiose by almost all contemporary Bayesians, because winning it requires a probabilist to countenance "logical" (or, at least, a priori) probabilities. But, Bayesians are still interested in rationally reconstructing peoples intuitive responses to The Paradox.

Bayesian Approach, I

- E_1 is $R(a) \wedge B(a)$ (a is a black raven)
- E_2 is $\neg R(a) \wedge \neg B(a)$ (a is a non-black non-raven)
- H is $\forall x(R(x) \rightarrow B(x))$ (all ravens are black)

Bayesian Approach, I

- E_1 is $R(a) \wedge B(a)$ (a is a black raven)
- E_2 is $\neg R(a) \wedge \neg B(a)$ (a is a non-black non-raven)
- H is $\forall x(R(x) \rightarrow B(x))$ (all ravens are black)
- K_α denotes the background corpus of information which consists of our (current) best understanding of the actual world. *But we will assume that K_α does not contain any specific information about the particular object a whose properties are at issue*

Bayesian Approach, II

(B) The degree to which E_2 confirms H relative to K_α is less than (perhaps much less than) the degree to which E_1 confirms H relative to K_α. Or, more formally, this claim becomes (given our assumption about the confirmation relation)
$Pr(H \mid E_2 \ \& \ K_\alpha) < Pr(H \mid E_1 \ \& \ K_\alpha)$

Bayesian Approach, III

Of course, when Bayesians assert (B), they are not merely making some descriptive claim like: "some actual agent S's conditional credences happen to be such that (B) is true of them."

Bayesian Approach, III

Of course, when Bayesians assert (B), they are not merely making some descriptive claim like: "some actual agent S's conditional credences happen to be such that (B) is true of them." Rather, they are making the claim that it would be reasonable for an agent S (who resides in the actual world as we know it) to be such that (B) is true of their credences.

Bayesian Approach, III

Of course, when Bayesians assert (B), they are not merely making some descriptive claim like: "some actual agent S's conditional credences happen to be such that (B) is true of them." Rather, they are making the claim that it would be reasonable for an agent S (who resides in the actual world as we know it) to be such that (B) is true of their credences. The idea here is that if it is reasonable to have (B)-like credences, then this can explain why it is reasonable to think there is something odd (if not paradoxical) about (PC). *(pg. 12)*