**"Inference to the Best Explanation, Cleaned Up and Made Respectable" - Jonah Schupbach**

Handout by Brett Park

The Charge: IBE, despite its apparent ubiquity in human reasoning, lacks a clear articulation:

> "As long as [IBE] is left vague, it seems to fit much rational activity. But when we scrutinize its credentials, we find it seriously wanting." - Bas van Fraassen, *Laws and Symmetry* p. 131

Schupbach wants to rectify this situation by giving IBE the probabilistic treatment

**Section 1**: Introduces a probabilistic notion of "explanatory power" and uses this to define a version of IBE, $IBE_p$

**Section 2.1**: Argues $IBE_p$ is cogent

**Section 2.2**: Computer simulations to check $IBE_p$ against competitors (Bayesian inference and chance)

**Section 1**: The formal account

Disclaimer: Any IBE argument appeals to a notion of "explanatory goodness" to select one explanation as the best. However, there are many different senses of 'explanatory goodness' (simplicity, unification, generality, power) Thus, there may be no general account of IBE.

Schupbach is concerned only with $IBE_p$, where "explanatory goodness" is "explanatory power" which he attributes to C.S. Pierce:

> The surprising fact, C, is observed;
>
> But if A were true, C would be a matter of course;
>
> Hence, there is reason to suspect that A is true.

Conditions we wish any measure of explanatory power $\mathcal{E}$ to satisfy:

> Condition 1: Positive explanatory power increases expectedness of the proposition
>
> Condition 2: Negative explanatory power decreases expectedness of the proposition
>
> Condition 3: No explanatory power does not increase or decrease expectedness of the proposition
>
> Condition 4: Maximum explanatory power leads to certainty of the proposition

Condition 5: Minimum explanatory power leads to certainty that the proposition of false

Condition 6: The more positive power over the proposition, the less positive power over its negation

First, we interpret expectedness as probability. Then we define a measure $\mathcal{E}(e,h)$ for explanatory power of hypothesis $h$ over evidence $e$ with the following desired mathematical structure:

$\mathcal{E}(e,h) \in [-1,1]$

$\mathcal{E}(e,h) = 1$ when a hypothesis $h$ has maximum explanatory power over evidence $e$

$\mathcal{E}(e,h) = -1$ when a hypothesis $h$ has minimum explanatory power over evidence $e$

$\mathcal{E}(e,h) = 0$ when a hypothesis $h$ has no power over evidence $e$

The additional formal adequacy conditions must be met:

**CA1 (Neutrality):** $\mathcal{E}(e,h) = 0$ if and only if $\Pr(h \wedge e) = \Pr(h) \times \Pr(e)$

**CA2 (Maximality):** $\mathcal{E}(e,h) = 1$ if and only if $\Pr(e|h) = 1$

**CA3 (Symmetry):** $\mathcal{E}(e,h) = -\mathcal{E}(\neg e,h)$

**CA4 (Irrelevant Conjunction):** If $\Pr(e \wedge h_2) = \Pr(e) \times \Pr(h_2)$ and $\Pr(h_1 \wedge h_2) = \Pr(h_1) \times \Pr(h_2)$ and $\Pr(e \wedge h_1 \wedge h_2) = \Pr(e \wedge h_1) \times \Pr(h_2)$, then $\mathcal{E}(e, h_1 \wedge h_2) = \mathcal{E}(e, h_1)$

From which we get (proof in the appendix):

**Theorem 1:** The only measure with a desirable mathematical structure that satisfies CA1–CA4 is

$$\mathcal{E}(e,h) = \frac{\Pr(h|e) - \Pr(h|\neg e)}{\Pr(h|e) + \Pr(h|\neg e)}$$

Example 1 (Symmetry): $\Pr(h|e) = .9$ and $\Pr(h|\neg e) = .5$

$$\mathcal{E}(e,h) = \frac{.9 - .5}{.9 + .5} = .29$$

$$-\mathcal{E}(\neg e,h) = -\frac{.5 - .9}{.5 + .9} = .29$$

Example 2 (Neutrality): $\Pr(h|e) = .9$ and $\Pr(h|\neg e) = .9$

$$\mathcal{E}(e,h) = \frac{.9 - .9}{.9 + .9} = 0$$

With $\mathcal{E}$, we can describe $\text{IBE}_p$ as the claim that hypothesis $h$ has greater explanatory power $\mathcal{E}$ than any of its competitors $h_i$. In argument form:

$e$

($\text{IBE}_p$) $\underline{\mathcal{E}(e,h) > \mathcal{E}(e,h_i), \text{ for any } h_i \text{ competing with } h}$

$\therefore h$

**Section 2.1**: $\text{IBE}_p$ and $\mathcal{E}(e,h)$ are cogent

(1) Positive $\mathcal{E}(e,h)$ always increase expectedness of $h$
(2) In the limiting case where our priors in $h \geq h_i$ are roughly equal, $\text{IBE}_p$ matches Bayesian inference

$$\frac{Pr(h \mid e) - Pr(h \mid \neg e)}{Pr(h \mid e) + Pr(h \mid \neg e)} > 0$$

$$\Leftrightarrow \Pr(h \mid e) > \Pr(h \mid \neg e)$$

$$\Leftrightarrow \frac{Pr(e \mid h)}{Pr(e)} > \frac{Pr(\neg e \mid h)}{Pr(\neg e)}$$

$$\Leftrightarrow Pr(e \mid h) - Pr(e \mid h)Pr(e) > Pr(e) - Pr(e \mid h)Pr(e)$$

$$\Leftrightarrow Pr(e \mid h) > Pr(e)$$

$$\Leftrightarrow Pr(e \mid h) > Pr(e \mid \neg h) \qquad \text{(L)}$$

$$\Leftrightarrow Pr(h \mid e) > Pr(h) \qquad \text{(C)}$$

When explanatory power is positive, our hypothesis always raises the probability of the evidence, and the evidence always raises the probability of our hypothesis.

This can help illuminate the "benign" circularity of IBE.

However, the central claim of $\text{IBE}_p$ is that a certain hypothesis $h$ has more explanatory power than any competitor $h_i$ ($h$ can be the best explanation if it is the least bad) so:

$$\frac{Pr(h\,|\,e)-Pr(h\,|\,\neg e)}{Pr(h\,|\,e)+Pr(h\,|\,\neg e)}>\frac{Pr(h_i\,|\,e)-Pr(h_i\,|\,\neg e)}{Pr(h_i\,|\,e)+Pr(h_i\,|\,\neg e)}$$

$$\Leftrightarrow \frac{Pr(h\,|\,e)}{Pr(h\,|\,\neg e)}>\frac{Pr(h_i\,|\,e)}{Pr(h_i\,|\,\neg e)}$$

$$\Leftrightarrow \frac{Pr(e\,|\,h)Pr(\neg e)}{Pr(\neg e\,|\,h)Pr(e)}>\frac{Pr(e\,|\,h_i)Pr(\neg e)}{Pr(\neg e\,|\,h_i)Pr(e)}$$

$$\Leftrightarrow Pr(e\,|\,h)-Pr(e\,|\,h)Pr(e\,|\,h_i)>Pr(e\,|\,h_i)-Pr(e\,|\,h)Pr(e\,|\,h_i)$$

$$\Leftrightarrow Pr(e\,|\,h)>Pr(e\,|\,h_i)$$

So, the hypothesis that is the most powerful explanation is the one that makes the evidence the most likely. This is equivalent to Bayesian inference:

$$\Pr(h|e) > \Pr(h_i|e)$$

$$\Leftrightarrow \frac{\Pr(h)\Pr(e|h)}{\Pr(e)} > \frac{\Pr(h_i)\Pr(e|h_i)}{\Pr(e)}$$

$$\Leftrightarrow \Pr(h)\Pr(e|h) > \Pr(h_i)\Pr(e|h_i)$$

except we are ignoring the prior probabilities of $h$ and $h_i$ when evaluating explanatory power.

*Assuming our hypotheses' prior probabilities are roughly equal*, we should conclude that $IBE_p$ does provide strong support for a hypothesis. $IBE_p$ is a "genuine epistemic virtue" (50).

**Section 2.2**: Computer trials

In this section, Schupbach uses computer simulated trials to compare the reliability of $IBE_p$ when compared to Bayesian approach and chance. The methodological steps are:

1. For each of a specified number $n$ of competing (mutually exclusive) explanatory hypotheses, assign values of the prior probabilities ($Pr(h_i)$) and likelihoods ($Pr(e|h_i)$). Priors and likelihoods are drawn randomly from a normal and uniform distribution, respectively (see discussion below for more details).
2. Using weights corresponding to the respective values of $Pr(h_i)$, randomly select the "true" hypothesis $h_j$ from $h_1, h_2, \ldots, h_n$. Each $h_i$ has a $Pr(h_i)$ chance of being selected.
3. Using the value of $Pr(e|h_j)$ (the likelihood associated with the true hypothesis), check whether $e$ "occurs." If $e$ occurs, continue with steps 4–6; otherwise, end this iteration.
4. Check which of the $n$ hypotheses has the greatest power; i.e., find $h_k$ where $\mathcal{E}(e,h_k) > \mathcal{E}(e,h_i)$ for all $i \neq k$.
5. Check which of the $n$ hypotheses is the most probable in light of $e$; i.e., find $h_l$ where $Pr(h_l|e) > Pr(h_i|e)$ for all $i \neq l$.
6. If $h_k = h_j$, count this as a case where the most explanatory hypothesis matches the true hypothesis; if $h_l = h_j$, count this as a case where the most probable hypothesis matches the true hypothesis.

Importantly, the simulation selects the true hypothesis proportionally to the prior probabilities. Thus, the priors match objective chance. The program is run for 2-10 hypotheses, each for a million iterations. Results:

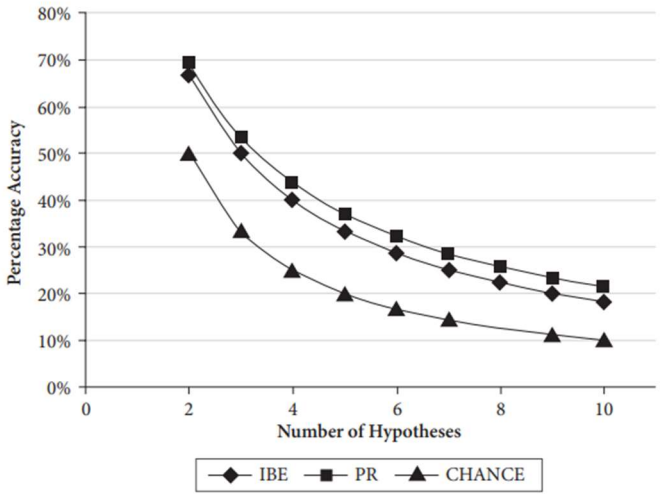


**Figure 4.1** Percentage accuracies in contexts that do not include a catch-all.

$IBE_p$ significantly outperforms chance and slightly underperforms Bayesian inference, regardless of the number of hypotheses or the catch-all. $IBE_p$ thus looks like "a poor man's Bayesianism" (54).

However, when we allow our priors to deviate from their objective values, the results change:

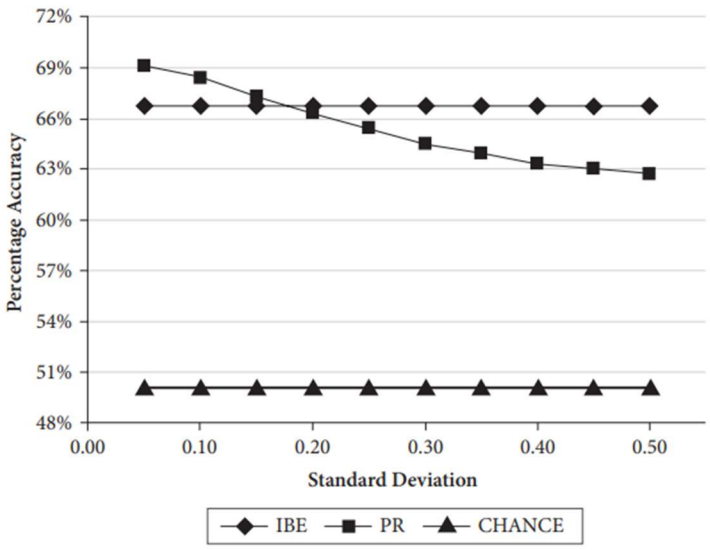


**Figure 4.3** Percentage accuracies in contexts that do not include a catch-all.

Which shows that the greater the deviations of our priors from objective chance, the more $IBE_p$ should be favored over Bayesian inference.

## Conclusions

- $IBE_p$ offers a precise, formal account of a IBE in terms of explanatory power $\mathcal{E}(e,h)$
- $IBE_p$ is cogent and reliable
- $IBE_p$ might be favorable over Bayesian inference when our priors have high variance

## A Worry

Schupbach cleans up "best" but not "explanation." The same counterexample that sank the DN model (as well as probabilistic theories of causation) appears to sink $\mathcal{E}(e,h)$ as well. Let $e$ be the proposition of a flagpole being a certain height and $h$ be the proposition of its shadow being the appropriate length at the appropriate time of day. $\Pr(h|e)$ is much higher than $\Pr(h|\neg e)$. By $\mathcal{E}(e,h)$, $h$ has large explanatory power over $e$. However, $h$ should have no explanatory power over $e$, because shadow lengths do not explain flagpole heights.