Overconfidence in Overconfidence

Kevin Dorst
kevindorst@pitt.edu

Formal Epistemology Workshop
23 May 2020

I. The Question

Do people tend to be overconfident in their opinions—i.e. more confident than they should be, given their evidence?

You are *calibrated* at confidence X% iff: of all the claims you are X%-confident in, X% are true.

Over-calibrated: less than X% true.
Under-calibrated: more than X% true.

Calibration studies: elicit confidence ratings; plot against proportion of truths. Result: "overconfidence effect".

For thresholds above 50%, people are often over-calibrated.

Psychologists infer that people tend to be overconfident. Call this inference—from "you are (mis)calibrated)" to "you are (ir)rational"—the **calibration inference**.

Summaries: Lichtenstein et al. 1982; Hoffrage 2004; Glaser and Weber 2010.

The Question: When and why is the calibration inference warranted? What does that tell us about the empirical results?

The Claims: When you should *defer* to a given person's evidence, the calibration inference is warranted. Otherwise, it's often not.

⇒ *Mis*calibration often evidence for *rationality*; the "overconfidence effect" is to be expected from rational people.

The Plan: problem (§II); insight (§III); limits (§IV); implications (§V).

II. The Problem

(1) (Average) actual confidence (credence) in some opinions.
(2) Proportion of those opinions that are true.
(3) (Average) *rational* confidence to have in those opinions.

Calibration inference: since (1) > (2), it must be that (1) > (3). This makes sense only if we should expect rational confidence to align with proportion true.

⇒ Requires explaining the connection between rationality and truth.

No necessary connection. Case 1: rational brains-in-vats.

Case 2: 60%-heads-biased coin I tossed 10 times yesterday. Credence in heads on each toss?

Case 3: Urn with 60 double-headed and 40 double-tailed coins; I'll draw one and tossed it 10 times. Credence in heads on each toss?

Upshot: We need a theory of when the calibration inference does (not) work.

III. The Insight

Simple scenario: Calvin took a test with random questions. Call all the claims he was 80%-confident in his **80%-opinions**. All you learn is how many of these were true.

Claim: in this simple scenario, the calibration inference works.

Parable: Magic Mary and her magic coins; coin-markings and toss-outcomes on stone tablets; bias-busting Bianca claims she can decipher them. How to test this claim?

Run a calibration test.

The **bias inference** works: we can infer from "Bianca is (mis)calibrated" to "Bianca can(not) decipher the coins."

If 79% of her 80%-opinions were true, good evidence that she can decipher the coin-markings; if 61% were true, good evidence that she can't.

Why? Because biases set expectations. You should *defer* to the biases of the coins in a robust (*independent*) way.

The analogy is between the **bias** of the coin for each tablet Bianca faces, and the **rational credence** given Calvin's evidence for each question he faces. When does rational credence set expectations?

What do *deference* and *independence* come to?

Let $q_1, ..., q_n$ be Calvin's 80%-opinions.

$q_i :=$ *the ith claim that Calvin was* 80%*-confident in (whatever it is) is true.*

Let $\overline{R}$ be the average rational credence for Calvin to have in the q_i. We want to know whether his 80%-opinions are (on average) rational ($\overline{R} = 0.8$), overconfident ($\overline{R} < 0.8$), or underconfident ($\overline{R} > 0.8$).

Where $R(q_i)$ is the rational credence for Calvin to have in q_i, $\overline{R} := \sum \frac{R(q_i)}{n}$.

Let P be a probability function representing *your* rational degrees of confidence.

Need $\overline{R}$ to set P's expectations.

> **Deference:** Upon learning that the average rational confidence for Calvin to have in his 80%-opinions is x%, you should be x% confident in each of them. $P(q_i|\overline{R} = x) = x.$

For all q_i, x.

Plausible in our simple scenario, since you have little/no evidence about q_i, and you know that Calvin has more.

Deference follows from 2 principles:
Point-wise Deference: For all q_i, and letting 'δ' be rigid: $P(q_i|R = \delta) = \delta(q_i)$.
Equality: For all q_i, q_j : $P(q_i|\overline{R} = x) = P(q_j|\overline{R} = x)$.

> **Independence:** Given that the average rational confidence for Calvin to have in his 80%-opinions is x%, learning that certain of these opinions are true or false shouldn't affect your opinion in the others. $P(q_{i_0}|\overline{R} = x, q_{i_1}, ..., q_{i_l}, \neg q_{i_{l+1}}, ..., \neg q_{i_k}) = P(q_{i_0}|\overline{R} = x).$

For all $q_{i_0}, ..., q_{i_k}, x$.

Plausibly approximately true in our simple scenario.

Exchangeability better; similar argument would go through.

Fact. Given Deference and (approximate) Independence, learning that x% of Calvin's 80%-opinions were true is good evidence that the average rational confidence for him is roughly x%.

$P(\overline{R} = x|\sum \frac{I(q_i)}{n} \approx x) > P(\overline{R} = x)$, where $I(q_i)$ is indicator of q_i.

> You think that *if* he's overconfident, each q_i is (say) 70% likely—so (since they're independent) you're confident that roughly 70% will be true. Likewise, you think that if he's rational, roughly 80% will be true. Etc. So the evidence you received (70% true) was much more likely on the overconfidence hypothesis than on the rationality or underconfidence hypotheses.

Precisely: $P(\sum I(q_i) \mid \overline{R} = x)$ is binomial with parameters x, n, and so has most of its mass centered around $x \cdot n$.

Example: $n = 50$; uniform over $\overline{R} = 0.6, ..., 0.99$.
$\Rightarrow$ If learn 70% of q_i true, credence that $\overline{R} < 0.75$ jumps from 37.5% to 78%.

IV. The Limits

By parallel reasoning, when Deference *fails*, the calibration inference will be inverted. For example, if conditional on his 80% opinions being rational you should be 70% confident in each, then learning that 70% of them are true is evidence that he's *rational*.

Precisely: if $P(q_i|\overline{R} = 0.8) = 0.7$, then $P(\overline{R} = 0.8|\sum \frac{I(q_i)}{n} = 0.7) > P(\overline{R} = 0.8)$.

> **Deference in Key:** Given (approximate) Independence, the the calibration inference works when and only when Deference holds.

Learning Calvin's **hit rate**—the proportion of all his answers that are true—breaks Deference and inverts the calibration inference.

Obvious in extreme cases; true more generally.

$P(q_i \mid \overline{R} = 0.8) = 0.8$, but
$P((q_i \mid \overline{R} = 0.8,\ H = 0) = 0$, and
$P((q_i \mid \overline{R} = 0.8,\ H = 0.1) \approx 0.2$

Given that his hit rate is low and that the (avg.) rational credence is x, you should be *less* than x-confident in each q_i.

$P(q_i|\overline{R} = x,\ H = 0.5) < x$, where $\boldsymbol{H}$ is overall hit rate on test.

$\rightarrow$ Rational credence *off* sets expectations: think that if he's rational, 70% of 80%-opinions true; if he's overconfident, 60% true; etc.

Assume that learning that Calvin's hit rate doesn't have strong effect on opinions about rationality: $P(\overline{R} = t|H = s) \approx P(\overline{R} = t)$.

This follows if we should expect Calvin to guess rationally—I'm skipping over an argument that we should.

Then if you know the hit rate is low (high), the calibration inference will often be inverted.

Example: $n = 50$, uniform over $\overline{R} = 0.6, ..., 0.99$; know $H = 0.5$; suppose $P(q_i|\overline{R} = x, H = 0.5) = x - 0.1$.
$\Rightarrow$ If learn 70% of q_i true, your credence that $\overline{R} < 0.75$ *drops* from 37.5% to 22%, and credence that $0.75 \leq \overline{R} \leq 0.85$ *jumps* from 27.5% to 61%.

V. The Implications

Claim: This epistemological story sheds doubt on the standard interpretation of the "fundamental bias in general-knowledge calibration" (Koehler et al. 2002, 687).

This is the **hard-easy effect**: on tests that have a hit rate below 75% (hard tests), people tend to be over-calibrated; on those with hit rates above 75% (easy tests), they tend to be under-calibrated.

The hard-easy effect subsumes the "overconfidence effect."

Standard explanation: people fail to account for difficulty of task.

Q: When a study that is hard (easy), what should we expect *rational* calibration curves to look like?

Claim: expect rational people to be miscalibrated.

We can simulate this. Suppose Bianca *can* decipher the biases of the coins. Two models:

> **Perfection model:** Always aligns credence exactly with bias.
>
> **Noise model:** Makes random (normally distributed) errors in aligning credence with bias.

What should we expect her (and by analogy: rational Calvin's) calibration curve to look like, given various hit rates?

Simulation: Random questions.

Others in paper; can discuss.

> *Result:* Hard-easy effect expected; realistic one on noise model.
>
> *Explanation:* Slack between bias and proportion true. If hit rate is low, two contributing factors: (1) fewer than normal extreme-bias coins; (2) fewer than normal coins landed the way they were biased. Bianca can account for (1), but not for (2).

Likewise for Calvin: tests with abnormally low (high) hit rates tend to be ones where rational opinions are over-(under-)calibrated.

What happens if we run this model on *my* test?

Upshot: We should expect the hard-easy effect to emerge for rational Bayesians.

VII. The Conclusion

The calibration inference is theoretically sound, but only under highly controlled conditions. Real studies don't meet them.

So to test the overconfidence hypothesis, we must predict the expected rational *deviations* from calibration on our tests. Doing so may overturn the standard interpretations of empirical effects.

References

Glaser, Markus and Weber, Martin, 2010. 'Overconfidence'. *Behavioral finance: Investors, corporations, and markets*, 241–258.

Hoffrage, Ulrich, 2004. 'Overconfidence'. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, 235–254.

Koehler, Derek J, Brenner, Lyle, and Griffin, Dale, 2002. 'The calibration of expert judgment: Heuristics and biases beyond the laboratory'. *Heuristics and biases: The psychology of intuitive judgment*, 686–715.

Lichtenstein, Sarah, Fischhoff, Baruch, and Phillips, Lawrence D., 1982. 'Calibration of probabilities: The state of the art to 1980'. In Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under Uncertainty*, 306–334. Cambridge University Press.