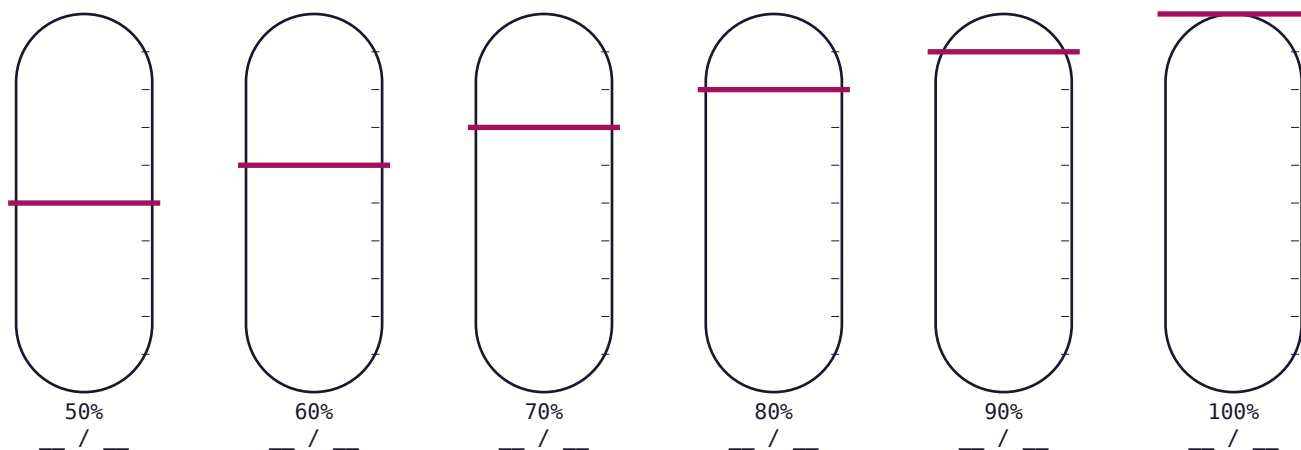


Uncertainty Laboratory

A rain-gauge observatory for calibrated confidence. A field guide to probability, calibration and intellectual humility, with a printable version of the calibration experiment, reflection prompts and verified references. By Sahil Tayal.

Labels used throughout. ESTABLISHED: replicated or definitional. CONTESTED: credible researchers disagree. HYPOTHESIS: plausible interpretation, not directly tested. ILLUSTRATIVE: simulation, invented example or unvalidated score — not data about people.



Reading a gauge: the magenta line is the confidence you claimed; shade the tube up to how often those claims came true. Liquid below the line = claims outran results; above = you knew more than you said.

1 • A vocabulary for uncertainty

Probability is a number between 0 and 1 attached to a claim. On a frequency reading it describes long-run rates; on a subjective (Bayesian) reading it is a degree of belief given evidence. Calibration bridges them: among all claims you rated 70%, about 70% should come true. Resolution asks whether your confidence moves with reality — whether your 90s beat your 60s; always forecasting the base rate is calibrated but useless. A proper scoring rule (Brier, logarithmic) makes honest reporting your best policy (Gneiting & Raftery, 2007). Moore and Healy (2008) separate three overconfidences: overestimation, overplacement and overprecision — the last the most persistent. Intellectual humility: in philosophy, proper attentiveness to and owning of one's intellectual limitations (Whitcomb et al., 2017); in psychology, recognising that one's beliefs might be wrong (Leary et al., 2017). Definitions still vary (Porter et al., 2022).

2 • Where miscalibration comes from

Feedback. Lichtenstein, Fischhoff and Phillips (1982) found experienced weather forecasters well calibrated while most other groups were not, and training gains limited; fast, unambiguous feedback is a plausible reason (HYPOTHESIS). Hard-easy effect. Overconfidence grows with task difficulty; easy sets can produce underconfidence (ESTABLISHED pattern, causes CONTESTED). Overprecision. People act as though they can make finer distinctions in uncertainty than they can. Random error. Grouping noisy judgments by stated confidence makes extreme bins sag toward the middle, mimicking overconfidence (ESTABLISHED mechanism; its share CONTESTED). Motivated confidence — preferring to feel sure — is plausible but not quantified by these sources (HYPOTHESIS).

3 • Two lenses on overconfidence

Lens A — a real bias. The heuristics-and-biases tradition documented confidence exceeding accuracy across tasks; Moore and Healy (2008) refined it; Mellers et al. (2014) trained against it. Lens B — partly method. Gigerenzer, Hoffrage and Kleinbölting (1991) argued that unrepresentative, tricky item selection produces apparent overconfidence; Erev, Wallsten and Budescu (1994) showed random error can produce over- and underconfidence in the same data depending on analysis. Brenner (2000) countered that error models do not make observed overconfidence spurious. A reasonable synthesis — editorial, not consensus — is that some measured overconfidence is real, some is produced by item sampling and analysis, and the mix varies by task (CONTESTED). This instrument is itself exposed to Lens B: its questions were chosen by one author and lean tricky.

4A • Case file: the forecasters (USA, 1950-1977)

Glenn Brier, a statistician at the U.S. Weather Bureau, proposed his score in a three-page 1950 paper in *Monthly Weather Review*. Murphy and Winkler (1977) summarised U.S. operational and experimental programmes and concluded that forecasters could formulate reliable subjective probability forecasts of precipitation and temperature (ESTABLISHED). Why it may have worked — daily repetition, next-day outcomes, honest scoring, statistical guidance — is interpretation (HYPOTHESIS). Limits: it does not show calibration transfers to one-off judgments, and calibration alone is not skill.

4B • Case file: “a serious possibility” (Washington, March 1951)

National Intelligence Estimate 29-51 called a Soviet-bloc attack on Yugoslavia in 1951 a “serious possibility”. In his 1964 essay *Words of Estimative Probability* (CIA Center for the Study of Intelligence), Sherman Kent recounted that he meant odds of about 65 to 35, a State Department planner had read much lower odds, and Board members who agreed the wording had privately

meant roughly 20% to 80% (PRIMARY HISTORICAL SOURCE). Kent proposed a chart (below) and a rule that unmodified “possible” carries no odds. He described resistance from “poets” who doubted words could carry precise odds — a genuine counterargument (CONTESTED). Limits: one episode, told by one participant thirteen years later.

Words	Probability	Give or take
Certain	100%	—
Almost certain	93%	about 6%
Probable (also “likely”)	75%	about 12%
Chances about even	50%	about 10%
Probably not	30%	about 10%
Almost certainly not	7%	about 5%
Impossible	0%	—

Kent's proposed chart, from his 1964 essay.

5 • Intellectual humility, properly sized

Calibration is humility you can count. The limitations-owning account places humility between intellectual arrogance (ignoring limitations) and intellectual servility (overweighting them). Leary et al. (2017) found self-rated humility correlated with openness, curiosity, tolerance of ambiguity and low dogmatism — correlational, self-report evidence. Underconfidence is a miss too: it wastes knowledge and cedes decisions to louder voices. Whether one can be “too humble” remains open (Porter et al., 2022).

6 • Counterarguments and ethical limits

Unique events have no frequency; numbers can create false precision; confidence may sometimes pay socially (open question); quiz calibration may not transfer to life; humility can be performed or unfairly demanded of the less powerful. Never use these scores to grade or select people. Never diagnose motives from a gap. Humility is not false balance — strong evidence deserves high confidence. Doubt can also be manufactured to delay action.

7 • What this laboratory does not prove

The gauges summarise a small, author-chosen question set answered once; they measure no stable trait. A gap inside the luck band is not evidence either way. The Noise Machine shows error can mimic overconfidence, not how much real overconfidence is noise. The Phrase Prism compares you with one 1964 proposal, not with current professional usage. Nothing here shows humility causes better decisions. Scores are illustrative and unvalidated.

Printable experiment · the paper instrument

Write an answer and circle one confidence for each question. Then score with the key on the next page, tally each confidence level onto the gauges, and compute: hit rate per level = right ÷ answered. Brier score = average of (confidence – outcome)², where outcome is 1 if right, 0 if wrong; always answering 50% scores 0.25. With 16 questions, expect large swings from luck — roughly ±12 points on your overall hit rate (one standard error). ILLUSTRATIVE.

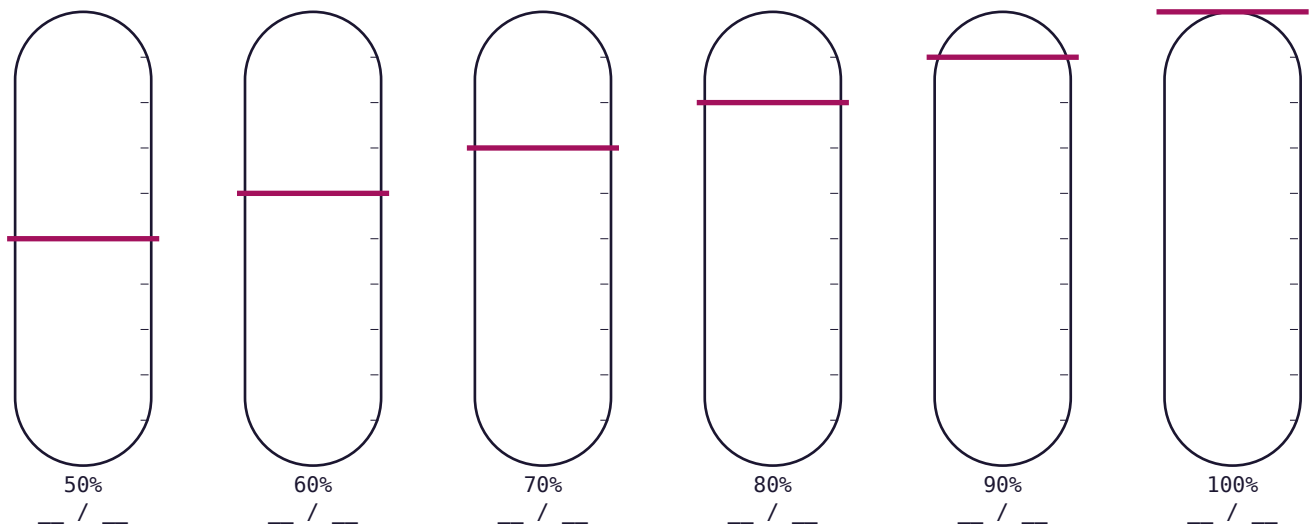
#	Question	Answer	Confidence
1	Which city lies farther north: London or Calgary?		50 60 70 80 90 100
2	Which is larger in area: Australia or Brazil?		50 60 70 80 90 100
3	Cleopatra lived closer in time to the building of the Great Pyramid, or to the Moon landing?		50 60 70 80 90 100
4	Which is heavier: a litre of water or a litre of olive oil?		50 60 70 80 90 100
5	Two fair dice: is a total of 7 or 8 more likely?		50 60 70 80 90 100
6	Patented first: the telephone, or an image-sending (facsimile) telegraph?		50 60 70 80 90 100
7	Hotter average surface: Mercury or Venus?		50 60 70 80 90 100
8	23 people: is the chance two share a birthday above or below 50%?		50 60 70 80 90 100
9	First: the first Nobel Prizes or the Wright brothers' first powered flight?		50 60 70 80 90 100
10	1% prevalence; test detects 90%, falsely flags 9%. After a positive, is the chance above or below 50%?		50 60 70 80 90 100
11	Which lies farther north: Rome or New York City?		50 60 70 80 90 100
12	Whales' closest living relatives: sharks or hippopotamuses?		50 60 70 80 90 100
13	Farther west: Los Angeles or Reno, Nevada?		50 60 70 80 90 100
14	More chromosomes per body cell: a human or a cultivated potato?		50 60 70 80 90 100
15	Larger: the seconds in a week, or one million?		50 60 70 80 90 100
16	Farther south: Detroit or Windsor, Ontario?		50 60 70 80 90 100

Answer key

1. London ($\approx 51.5^\circ\text{N}$ vs 51.0°N)
2. Brazil (≈ 8.5 vs 7.7 million km^2)
3. The Moon landing
4. Water (olive oil ≈ 0.91 kg/L)
5. 7 ($6/36$ vs $5/36$)
6. The facsimile telegraph (Bain 1843; Bell 1876)
7. Venus
8. Above ($\approx 50.7\%$)
9. Nobel Prizes (1901 vs 1903)
10. Below ($\approx 9\%$)
11. Rome ($\approx 41.9^\circ\text{N}$ vs 40.7°N)
12. Hippopotamuses
13. Reno ($\approx 119.8^\circ\text{W}$ vs 118.2°W)
14. Potato (48 vs 46)
15. One million (a week = 604,800 s)
16. Windsor

Your gauges

Under each tube write right / answered. Shade the tube to that proportion. Compare the shading with the magenta line.



The Noise Machine, on paper

In the website's fictional simulation, a judge with zero bias but noisy reports looks overconfident when its judgments are grouped by stated confidence: the top bins collect lucky overshoots. A perfectly calibrated judge with noisy evidence looks calibrated by stated confidence but underconfident when grouped by true probability. This is the symmetry Erev, Wallsten and Budescu (1994) described; how much real overconfidence it explains is contested (Brenner, 2000).
FICTIONAL SIMULATION.

The Phrase Prism, on paper

Before reading Kent's chart, write the percentage you mean by: almost certain ____ · probable ____ · likely ____ · chances about even ____ · probably not ____ · almost certainly not ____ · a serious possibility ____ · possible ____ . Ask one other person to do the same. Compare. Where do you disagree by more than 20 points?

The Forecast Diary — reflection prompts

01 Write one belief you hold at 90% or more. What would you expect to see if it were false?

02 When you say "probably", what number do you mean? What number does someone close to you hear?

03 Recall a time you were certain and wrong. What feedback was missing, slow or ignored?

04 Choose one decision this week. Write a probability for its key outcome and a date to check.

05 Where might you be underconfident? What does that cost, and whom?

06 In one current disagreement, what would intellectual humility look like without servility?

References (verified 27 October 2026)

1. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
2. Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12, 595–600.
3. Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *JRSS Series C*, 26(1), 41–47. doi:10.2307/2346866
4. Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In Kahneman, Slovic & Tversky (Eds.), *Judgment under Uncertainty* (pp. 306–334). Cambridge University Press.
5. Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models. *Psychological Review*, 98(4), 506–528.
6. Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence. *Psychological Review*, 101(3), 519–527.
7. Brenner, L. (2000). Should observed overconfidence be dismissed as a statistical artifact? *Psychological Review*, 107(4), 943–946.
8. Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2), 502–517. doi:10.1037/0033-295X.115.2.502
9. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *JASA*, 102(477), 359–378. doi:10.1198/016214506000001437
10. Tetlock, P. E. (2005). *Expert Political Judgment*. Princeton University Press.
11. Mellers, B., et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5), 1106–1115. doi:10.1177/0956797614524255
12. Hauenstein, C. E., Thomas, R. P., Illingworth, D. A., & Dougherty, M. R. (2025). Rethinking the role of teams and training in geopolitical forecasting. *Psychological Science*, 36(1). doi:10.1177/09567976241266481
13. Kent, S. (1964). Words of estimative probability. CIA Center for the Study of Intelligence. cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf
14. Whitcomb, D., Battaly, H., Baehr, J., & Howard-Snyder, D. (2017). Intellectual humility: Owning our limitations. *Philosophy and Phenomenological Research*, 94(3), 509–539. doi:10.1111/phpr.12228
15. Leary, M. R., et al. (2017). Cognitive and interpersonal features of intellectual humility. *PSPB*, 43(6), 793–813. doi:10.1177/0146167217697695
16. Porter, T., et al. (2022). Predictors and consequences of intellectual humility. *Nature Reviews Psychology*, 1, 524–536. doi:10.1038/s44159-022-00081-9

Sahilverse

Uncertainty Laboratory is an independent educational project by Sahil Tayal, part of the Sahilverse (sahiltayal.com). Not academic research, a psychometric test or professional advice. Corrections and better sources: sahil@sahiltayal.com. Related worlds: risks.sahiltayal.com · decisions.sahiltayal.online · thegoodscientist.sahiltayal.com · bias.sahiltayal.online