Scoring rules and expert disagreement

The problem

Ade is a policymaker, trying to decide how to enhance Thames flood defences for the next fifty years. He wishes to use the best scientific advice available to determine the likelihood that the Thames will rise more than 50cm—which would require new barriers. He convenes a panel of experts. The 10 experts disagree, offering a wide range of answers from unlikely to very likely.

What is the rational response for a layperson in the face of expert disagreement?

- Ade does not understand the evidence or the dispute—cannot judge for himself
- Any expert knows more than Ade—rational to *defer* to any one. But…

(Hardwig 1985, 1991; Goldman 2001)

PART I – MANY TO ONE

1 Common answer: aggregate

Background thought: Ade can't choose, adjudicate. The testimonial evidence clusters in a certain way. Can we use facts about the group's distribution to get closer to the truth?

Simplest case: equal-weight linear average
More common: weighted linear average
Other options: geometric average, median, proposition-based voting

Justification: Condorcet Jury Theorem, wisdom of crowds, statistical results on sampling
Note: **Weighting** is common to many approaches, and most used in cases I have examined

(Cooke and Goossens 2000; Dietrich and List 2016)

2 Expert elicitation and weighing expertise

Use track records of successful prediction of this type. Real or tested.

Common method:

- Determine relevant domain, design test questions that are expected to match target question. Test skill, not memory: we know answer, experts work it out. Must be measurable, with clear answers. (Work with experts here)
- Test all experts, score results. Relative scores become *weights*
- Conduct elicitation, adjust each answer by expert's weight, average

e.g., Weather forecasts.

(Cooke 1999; Aspinall 2010)

3 How do you score the test?

There was a 52% chance of rain this afternoon. How good was this prediction?

What if it had been: 30%, 70%. *How much* better/worse?

Error: event – probability	1 point when event & probability > 50%, add up points
Brier score	1 point to best forecast on each question
Chi-squared test	Weight points depending on importance of question

There are lots of these! >50 for kinds of weather forecasts

How to choose? Epistemic properties

- Calibration: how closely the expert's opinion matches the data. e.g., statement assessed with probability of X% was true X% of the time, for the observed X
- Entropy: A measure of how "spread out" probabilistic judgements are. Lower entropy is favoured.
- Propriety: Expert receives maximal score iff they state their true opinion. No benefit to adjusting.

(Australian Bureau of Meteorology 2017; Cooke 1999)

4 Good news and bad news

Bad: disagreement over scoring rules is a case of expert disagreement! Ade has attempted to resolve one such problem and arrived at another.

Good: lots of different expert disagreement problems collapse to one. If we can resolve how to do expert elicitation and aggregation, we have made much progress!

Recommendation 1.0*: Policymakers should invest in building expertise in this one domain (expert elicitation), as all other expert disagreement problems reduce to it. (Assuming the justification of aggregation goes through!)*

PART II – HIGHER-ORDER INDUCTIVE RISK

BUT: There is more to scoring rule selection than settling on epistemic values and optimising for them. *Scoring rules encode non-epistemic values.*

5 Background: hypothesis testing and values

Binary forecasts (e.g., Positive/Negative) → how certain should we be to make a forecast?

e.g., Pregnancy test. Indicator detects pregnancy X% of the time. How high should X be? How many false positives/negatives are acceptable?

	Observed Positive	**Observed Negative**
Forecast Positive	True Positive	False Positive
Forecast Negative	False Negative	True Negative

Answer: it depends what you care about! Non-epistemic value question

Inductive risk and values in science

- Experimental results make certain hypotheses more/less likely
- Scientists accept hypotheses once reaches some threshold: sufficiently certain
- But how much is sufficient? "How sure we need to be before we accept a hypothesis will depend on how serious a mistake would be"
- *Non-epistemic values* determine the acceptance conditions for scientific hypotheses

(Rudner 1953; Douglas 2000)

6 Scoring rules and values

Test statistics are simple scoring rules:

Accuracy	TP+TN / Tot	Simple. Doesn't differentiate TP/TN
Sensitivity	TP / Obs P	"Hit rate." Sensitive to FNs. Ignores FPs
Bias	Fore P / Obs P	Over/underforecast. Relative freq only
…	…	… *(ABM lists 12 tests for binary forecasts)*

Which should we use? *Depends what we care about.*

Examples: suppose we need to aggregate multiple experts/tool for these purposes

- HIV test – avoid FN
- High investment decisions – avoid FP
- Sharing sweets – optimise for TPs

} Non-epistemic values do the work!

This simple case generalises. Even once we fix epistemic values, there are multiple rules. They differ in terms of "epistemic risks"—risks of different kinds of false beliefs—and the dispute can only be adjudicated by non-epistemic values.

(Moss 2011; Babic forthcoming)

7 Whose values?

They must be the policymakers.

- Ade will make judgements using the aggregate opinion
- To do so, he will apply his values to those probabilities
- But: different options for aggregations exist. They cater differently to different risks
- Claim: Ade can reasonably select the aggregation that best caters to the risks he cares about.
- His values can determine how the expert disagreement is resolved

Objection: (Higher-order) Value-Free Ideal

- Didn't Ade want the experts to tell him *what will happen*?
- By choosing a rule that conforms to his values, he is crafting the truth to fit his tastes

Reply

1. Not quite the same as the VFI in science. There, we worry about scientists bringing values into their inquiry. It might be that science can be value-free up to the point of aggregation, but *here* someone's values must enter.
2. Respect's Douglas's demand for values to play *indirect* role: they're for managing uncertainty, and constrained by evidence ← note disagreement = uncertainty

8 Final recommendation

Investing in expert elicitation and understanding scoring rules allows policymakers to navigate the problem of expert disagreement

Expert disagreement *isn't* a purely epistemic problem. Policymakers need clarity on the values they will apply, and how to apply them, to select a scoring rule

References

Aspinall, Willy. 2010. “A Route to More Tractable Expert Advice.” *Nature* 463 (21): 294–25.

Australian Bureau of Meteorology. 2017. “Forecast Verification,” 2017. http://www.cawcr.gov.au/projects/verification/.

Babic, Boris. forthcoming. “A Theory of Epistemic Risk.” *Philosophy of Science*. https://philpapers.org/rec/BABATO.

Cooke, Roger. 1999. *Experts in Uncertainty*. Oxford: Oxford University Press.

Cooke, Roger, and L. H. J Goossens. 2000. “Procedures Guide for Structured Expert Judgement in Accident Consequence Modelling.” *Radiation Projection Dosimetry* 90 (3): 303–9.

Dietrich, Franz, and Christian List. 2016. “Probabilistic Opinion Pooling.” In *Oxford Handbook of Probability and Philosophy*, edited by Alan Hajek and Christopher Hitchcock. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199607617.013.37.

Douglas, Heather. 2000. “Inductive Risk and Values in Science.” *Philosophy of Science* 67 (4): 559–79. https://doi.org/10.1086/392855.

Goldman, Alvin I. 2001. “Experts: Which Ones Should You Trust?” *Philosophy and Phenomenological Research* 63 (1): 85–110. https://doi.org/10.1111/j.1933-1592.2001.tb00093.x.

Hardwig, John. 1985. “Epistemic Dependence.” *Journal of Philosophy, Inc* 82 (7): 335–349.

———. 1991. “The Role of Trust in Knowledge.” *Journal of Philosophy* 88 (12): 693–708.

Moss, Sarah. 2011. “Scoring Rules and Epistemic Compromise.” *Mind* 120 (480): 1053–1069.

Rudner, Richard. 1953. “The Scientist Qua Scientist Makes Value Judgments.” *Philosophy of Science* 20 (1): 1–6. https://doi.org/10.1086/287231.