Making confident decisions with model ensembles

Joe Roussos

Richard Bradley

Roman Frigg

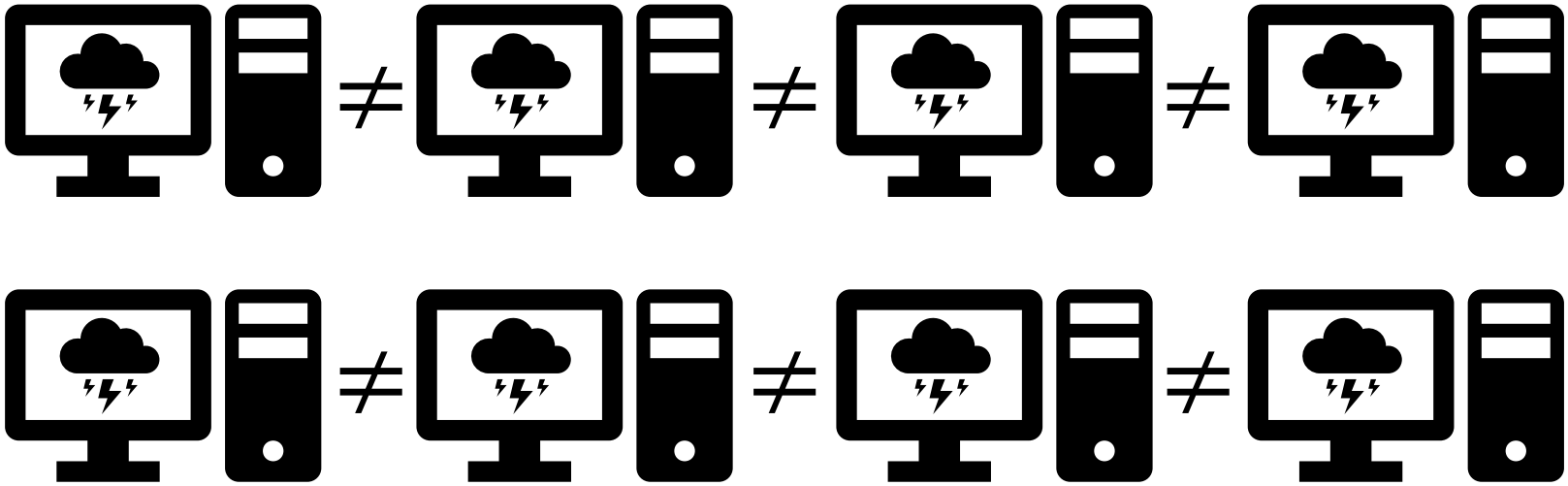
1 May 2019

CHESS, Durham

London School of Economics

The Problem

How do we use model ensembles to inform decision-making, in a way that reflects and makes use of scientific uncertainty?



Our Decision Problem

How much does it cost to offer insurance against natural catastrophes?

Basics of the Decision Problem

The ingredients of insurance pricing

(1) **Probability** that the insured event *E* happens

(2) **Pricing model**, which is a function of things we know about *E* including, critically, its probability

Important but often forgotten:

(3) Assessment of **uncertainty** about (1)

Toy Example

Simplified problem:

- You are a new insurer
- You want to sell a *single* insurance contract on house damage due to hurricanes.
- Event *E*: “a hurricane strikes Fort Lauderdale in 2020”.
- The contract is for a total value v=$100,000.
- Binary contract: paying out either $0 if the event does not occur, or $100,000 if it does.

What should you charge for this contract?

The Dream Answer

Take perfect model and calculate $p(E)$, and plug it into a pricing model.

But …

- There is no perfect model. Of necessity models omit factors (known and unknown) and make idealisations.
- Many models exist and impossible to decide between them on the basis of available evidence.
  - *Florida Commission on Hurricane Loss Prevention* 2007 assessment: ensemble of 972 models.

The Real Answer

- Buy an ensemble of predictive models from a commercial modelling company
  - The ensemble members are chosen such that they reflect scientific disagreement
  - There are known inadequacies with all models

→ How does this ensemble inform pricing?

- Today: model averaging
- …we think we can do better

Plan

1. Insurance Pricing: a primer
2. Confidence Approach, via the toy example
3. Elaborations
4. Questions

1. Insurance Pricing

What is the minimum price for an insurance contract covering event *E*?

Key point: **function of the probability** of *E*.

This is all you need to bear in mind if you are not into insurance …

... I will show a bit more now because the detail helps show why this is important.

Pricing profitably

Stone's (1973) constraint equation:

π : minimal annual premium required for profit

$$\pi > \begin{cases} \text{Expected payout on } E + \\ \text{Cost of holding the capital to insure } E \end{cases}$$

Example: Given your contract on *E*, you expect to pay out \$1k. To offer insurance on *E* regulations require you hold \$100k. This costs you \$5k. So you must charge at least \$6k for your insurance.

More Formally

Stone's (1973) constraint equation (simplified):

$$\pi > \langle d \rangle + Hy$$

$\langle d \rangle$: expected damage (pay out)

H : capital holdings required to insure *E*

y : cost of holding capital in %.

Calculating the components

$$\pi > \langle d \rangle + Hy$$

- Damages: $\langle d \rangle = \sum_E p(E) d(E)$
- Cost of Capital: y is an opportunity cost, we'll just think of it as a bank rate
- Holdings: how is H determined?
  - Layman's guess: the total cost of E-contract
  - In reality this would be too expensive!

Holdings

In practice, a regulator tells insurers how much of their book their need to be able to cover

We can ignore these complex details for now!

In a *single contract* scenario, the layman's guess is correct!

- Regulators would demand that the insurer holds the full amount

Let's go back to the Toy Example

- You are a new insurer
- You want to sell a *single* insurance contract on house damage due to hurricanes.
- Event *E*: "a hurricane strikes Fort Lauderdale in 2020".
- The contract is for a total value v=$100,000.
- Binary contract: paying out either $0 if the event does not occur, or $100,000 if it does.

What should you charge for this contract?

Toy example "model outputs"

You consult a modelling firm

"There are 10 models, and they disagree!"

Model	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
p(E)	0.0070	0.0083	0.0071	0.0074	0.0091	0.0076	0.0061	0.0092	0.0068	0.0086
Weight	**0.2368**	0.0729	0.2071	0.1575	0.0158	0.0317	0.1157	0.0173	0.1148	0.0300

"But we've assessed them for predictive skill…

…weighted each model for performance…

…and averaged them, using scoring rule R"

"The answer is *p(E)=0.0072*"

Pricing your contract

Using the ensemble average one finds:

$$\langle d \rangle = \$720$$

As only one event is insured we know

$$H = \$100{,}000$$

Let us assume

$$y = 5\%$$

So the result is:

$$\pi > \$5{,}720$$

Problems with averaging

1. Dataset for historical hurricanes is small, so evidential basis for predictive testing is small
2. Climate change means we expect future patterns may differ significantly from historical record, raising doubts about “hindcasting”
3. Models often contain contradictory assumptions, so it is unclear that averaging them is coherent
4. Great many scoring rules on offer, with experts disagreeing over which is best
5. Averaging discards useful information about the state of scientific uncertainty

Uncertainty "management"

Insurers tell us they are concerned about unaccounted for uncertainty in this process (we think: reasonably!)

They therefore introduce an "inflationary factor" for safety (we think: entirely arbitrarily)

Here we will represent this crudely by doubling the probability to yield the higher "safety" price.

With the same H and y one gets:

$$\pi > \$6{,}438$$

(this change is artificially *small* due to our toy setup)

2. The Confidence Approach

We want to make explicit use of the following:

- What is at stake in the decision
- Uncertainty attitude of the decision maker
- The nature and spread of evidence available

Point of departure: There is something wrong with the “linear” process of decision-making just described.

Issues of model uncertainty and their effect on the “answer” cannot be separated from the decision problem we want to solve.

Overview of the Approach 1

1. Assess **how important** the decision is to the agent
   - Compared to other decisions they make
   - Using whichever measure of importance they prefer
2. Using this, settle on **how confident** the agent wants to be to make this decision
   - Less important decisions require less confidence…
   - We will use simplified levels of confidence: Low, Medium, High
   - Confidence is generated by **weight of evidence**, a function of quantity, quality, adequacy, diversity, etc.

Overview of the Approach 2

3. Use the **scientific evidence** (model outputs) to construct a series of potential answers
    - We will consider a set of nested claims—representing less specific but more reliable potential answers
    - Each trades off specificity and robustness in different ways
    - We then classify them into the levels of confidence that they licence: Low, Medium, High
4. Finally, we **select** the claim which best fits the confidence that the decision-maker required

Preliminary notes

1. We **invert the "linear model"** of policy advice
   - Instead of the scientist supplying the policymaker with a prefabricated right answer,
   - the process begins with the decision-maker's needs
   - and the answer is built to fit those needs as best possible given the evidence
2. **"Weight of evidence" is complex** – this will be one of our links to adequacy
3. We expressly **build in the decision-maker's attitude to uncertainty** in the form of their demand for confidence
   - Our approach makes no claims about the superiority of one response to uncertainty

Step 1: What is at stake?

Stakes of the decision: the agent's assessment of how important it is.

e.g., What's the worst that could happen?

Convention: number on a 0-to-1 scale.

0: You don't care (e.g. £1 bet)

1: Highly significant (e.g. you're shot if you lose)

Toy example: Stakes

- This contract will constitute your whole business and so the risk of ruin is very high.
- Still, no one's life is at stake and there is no impact on anything outside of the realm of this decision.

Conclusion: The stake is moderately high

Let us use s=0.75

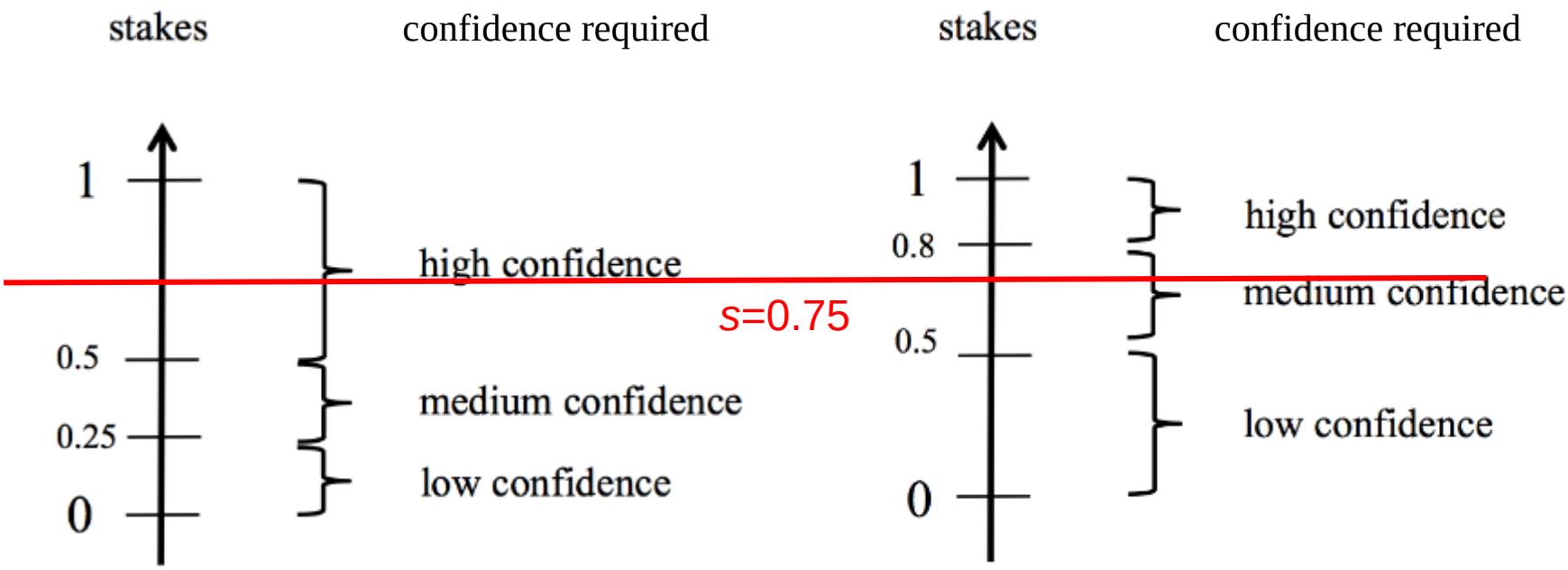
Step 2: Cautiousness

Given the importance of the decision, how confident do you want to be in order to act?

→ Cautiousness: function from stakes to “levels of confidence”

- Cautiousness represents uncertainty *attitude*.
- It will be subjective and will need to be elicited.
- Can be different for different agents.

Simple Examples: Cautiousness



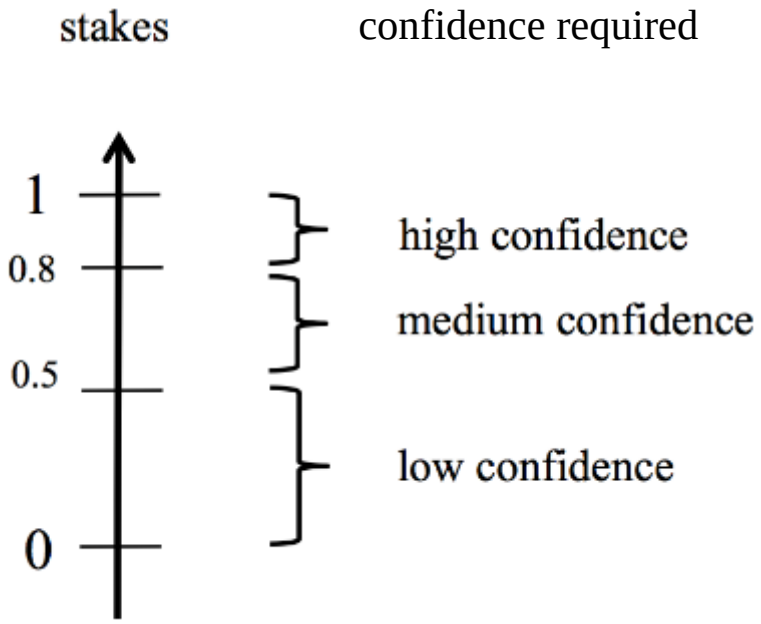
Cautious Agent

Bold Agent

Toy Example

Insurance of natural catastrophes involves significant uncertainty, so you can't be *overly* uncertainty averse.

Let's use the "bold" attitude:



Progress

We now know

(1) How important the decision-maker thinks this decision is: ***s=0.75***

(2) Given that, how confident they want to be in order to decide: ***Medium***

In real situations, (1) and (2) will both be informed by other decisions they make, and the nature of their field

Step 3: Nested Intervals

We now turn to the evidence base: in this case, outputs from models

We use these to construct a series of nested claims

The probability of a hurricane striking Fort Lauderdale in 2019 is…

- = 0.007
- between 0.007 and 0.0072
- between 0.0068 and 0.0072
- …

Where these values are **model outputs**

How to build the nested sets

Assumption: there is a best model, we know it

In our example: m_1 – *chosen by scoring rule R*

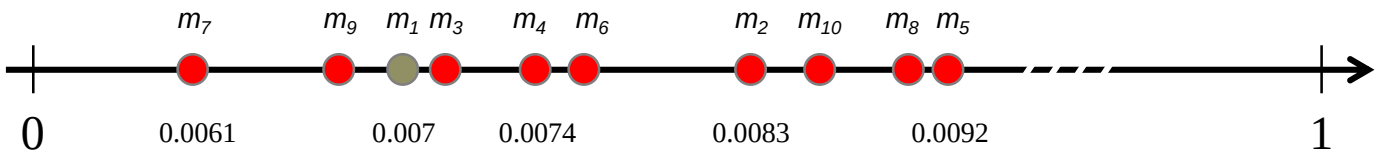
Our lowest level, most specific claim is that the probability of the event just is 0.007.

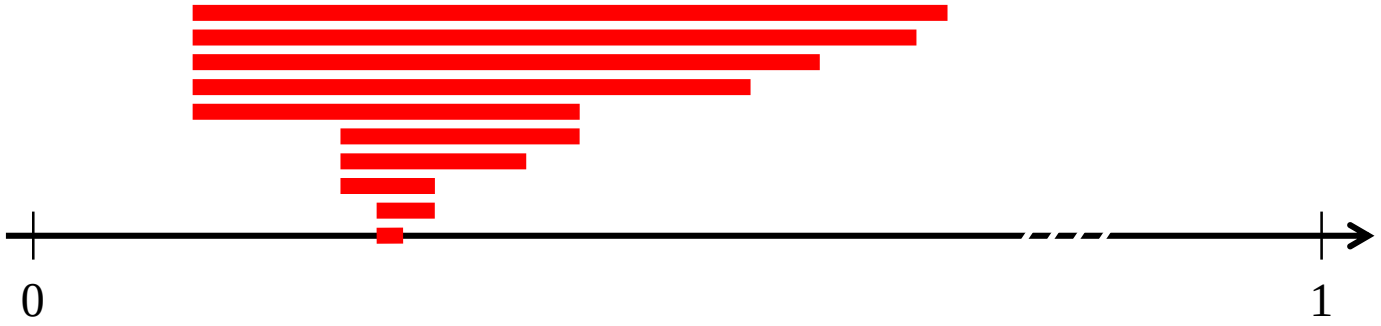
We can form wider intervals by including the predictions in the order of their distance from the best model.

We'll return to this in "Elaborations"

Step 3: Nested Intervals

From model outputs to nested intervals





Step 4: Confidence Grading

Confidence is generated by examining the weight of evidence supporting a claim.

Aim: attach to each interval a confidence level.

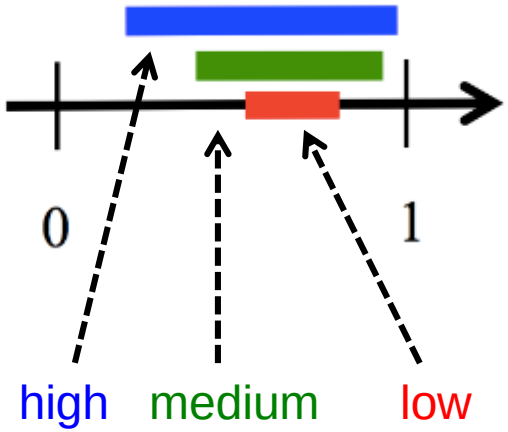
Important: increased confidence does *not* change the probabilities; it makes us more confident that probabilities we have are right.

Step 4: Confidence Grading

Convention:

- Three “levels” of confidence
- Low, Medium, High

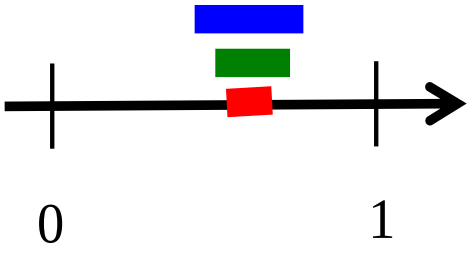
Logic dictates: less confidence in more precise claims.



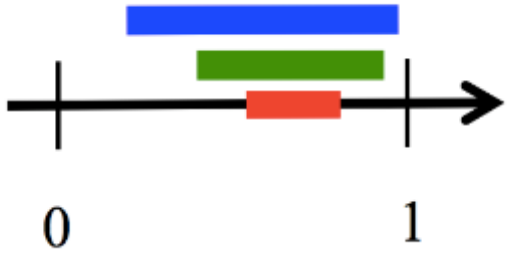
Step 4: Confidence Grading

Confidence grading reflects the state of scientific understanding

- Width of intervals reflects confidence-precision trade-off on some claim
- Wide/narrow intervals show that weight of evidence for projections is low/high.



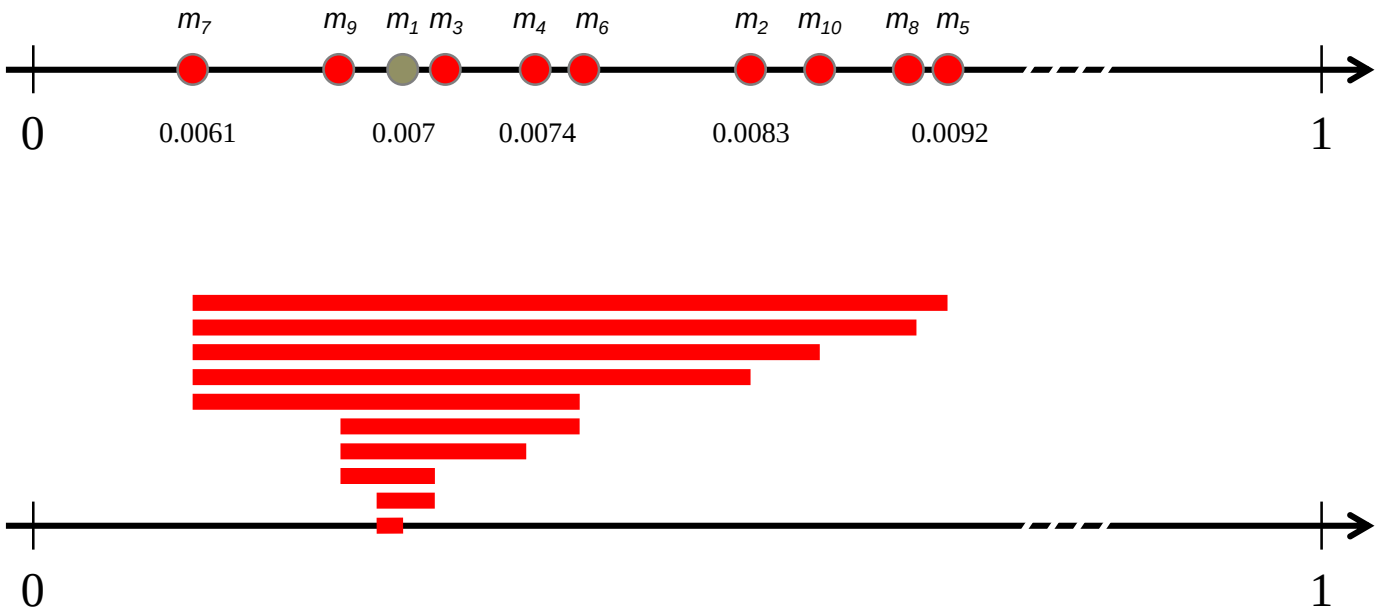
Good understanding



Poor understanding

Toy Example

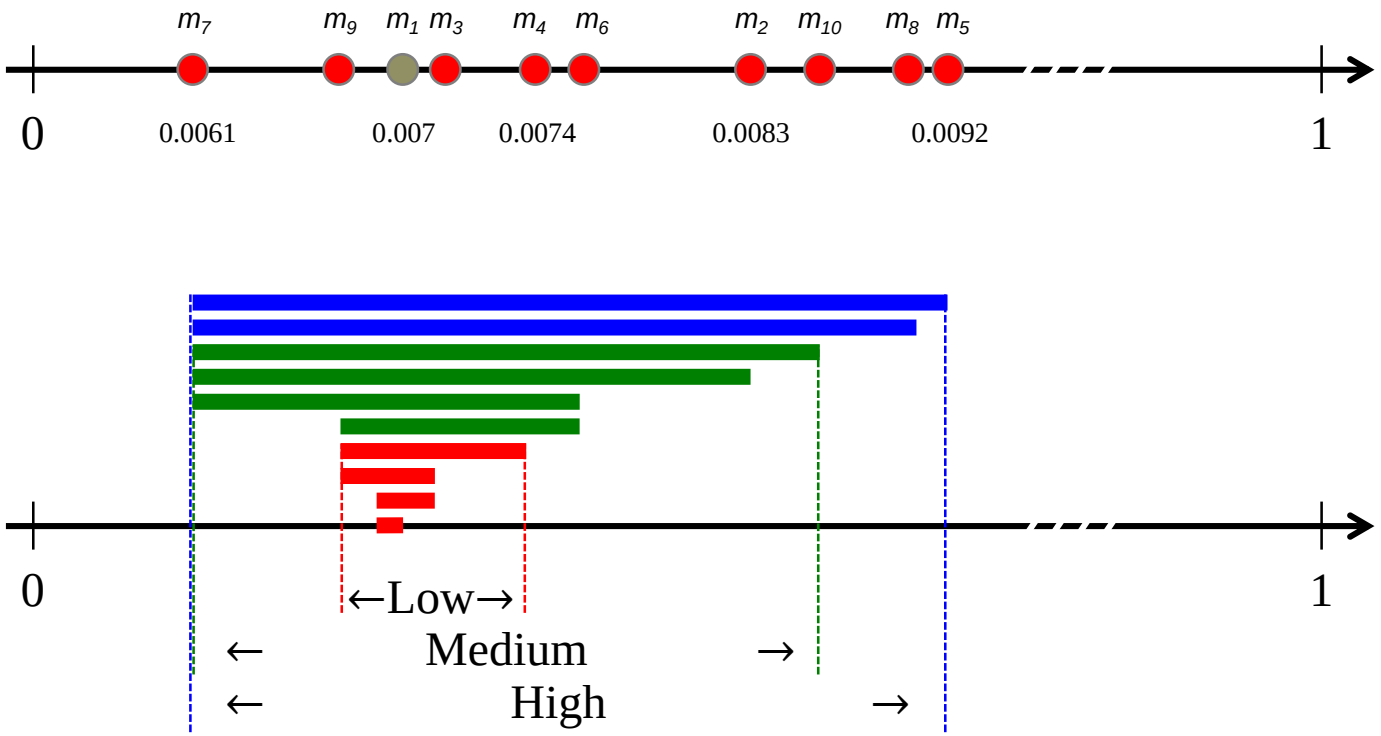
Recall our nested interval structure



The only evidence I've described is the collection of model outputs → they are all we can use to assign confidence

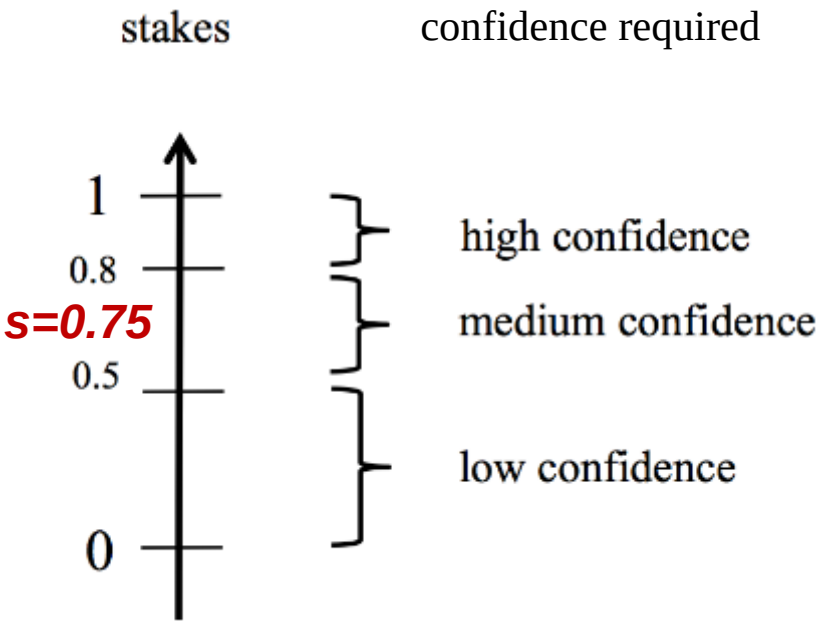
Toy Example

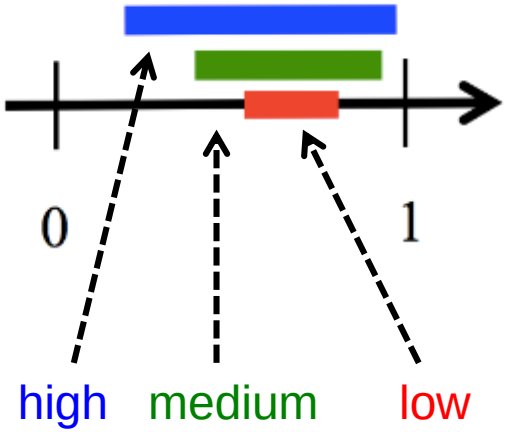
Step 4: Assign confidence levels



Step 5: Select “your” interval

Apply your cautiousness function to your assessed stakes

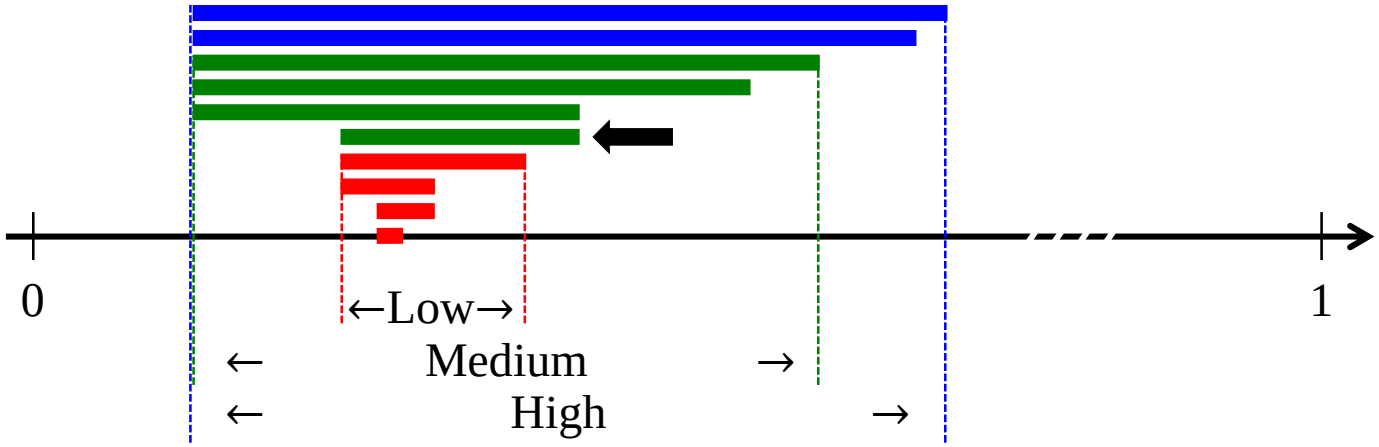




You choose the **green** (medium) interval.

Toy Example

Given s=0.75, **medium** confidence is required.



The narrowest interval in the medium range is I_5=[0.0068,0.0076].

Choose this interval for the decision!

Progress

We have settled on a particular interval of probabilities, I_5=[0.0068,0.0076].

It represents a particular trade-off between

- **specificity** (narrowness of the interval), which is valuable in distinguishing between courses of action, and
- **robustness** (breadth), which ensures we are confident enough in our decision given the uncertainty

Step 6: Make a decision

What we have now is a **set of probabilities**. We need a decision-rule that works with these. There are many candidates; we will use:

Maximin Expected Utility: *A* is preferred to *B* iff the minimum expected utility of *A* is greater than the minimum expected utility of *B*.

Colloquially: Choose the option that has the best outcome if things turn out to be as bad as they can be.

Toy Example

Our interval is I_5=[0.0068,0.0076].

Let's assume for simplicity that things go badly when the probability is highest

Therefore work with *p(E) = 0.0076*

Compare: averaging *p(E)=0.0072*

(We're 5% higher)

Pricing your contract

$$\pi > \langle d \rangle + Hy$$

Using the confidence-derived probability:

$$\langle d \rangle = \$760$$

As only one event is insured we still have

$$H = \$100{,}000$$

Let us again assume

$$y = 5\%$$

So the result is:

$$\pi > \$5{,}7\mathbf{6}0$$

Comparison

Averaging (simple)	$5,720
Averaging + “safety”	$6,438
Confidence	$5,760

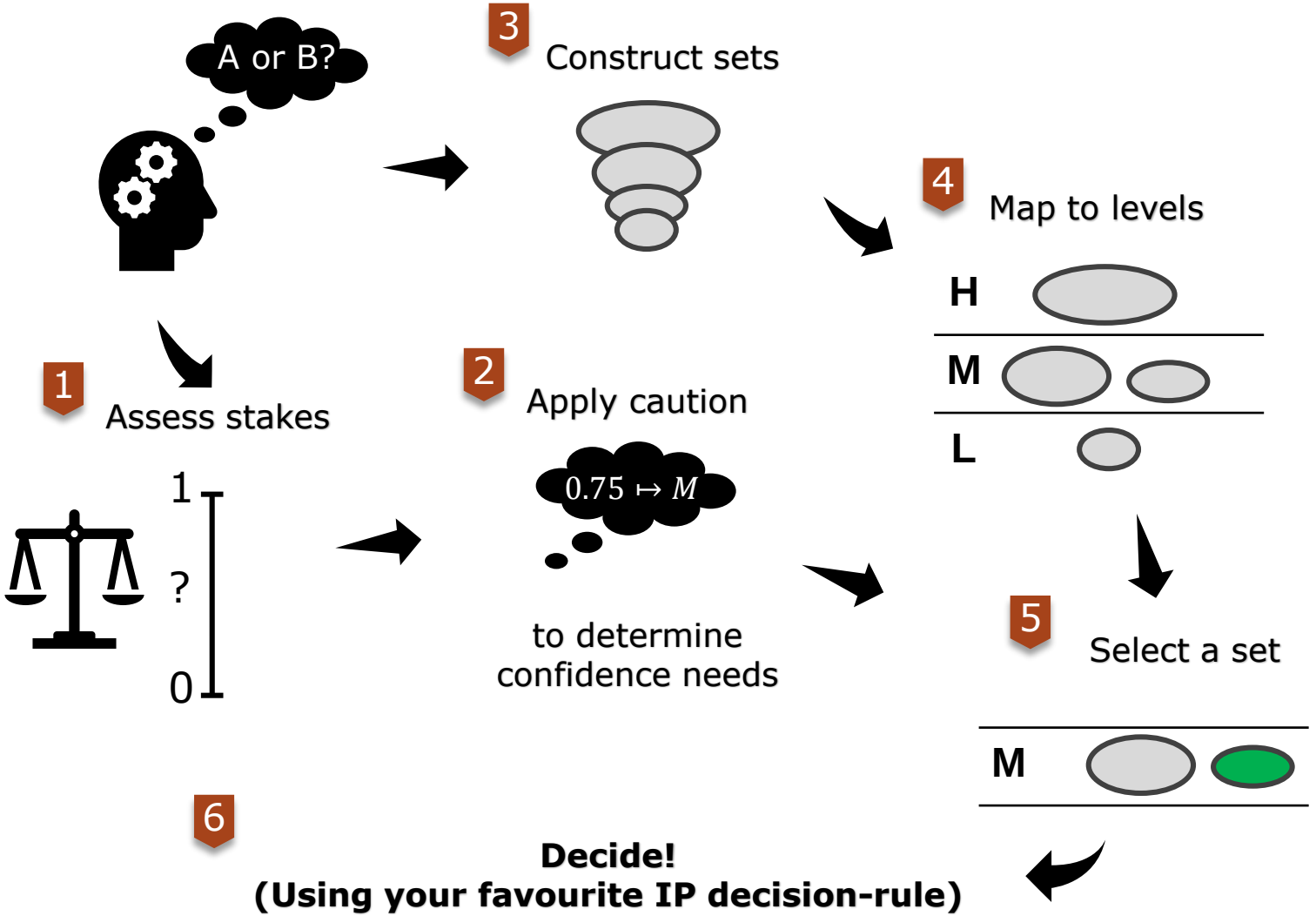
Note:

Confidence is 0.7% above Averaging,

but 10.5% smaller than the arbitrary “safety” price.

“Rule of thumb” uncertainty management is not only baseless, it is also not cost-effective.

Summary of the Procedure



3. Elaborations

The method just presented is idealised in key ways

- No discussion of the details of the models, esp. *why* they disagree
- Intervals formed by Euclidean distance, no facts about models used except where their outputs fall
- Still used a scoring rule (to pick the centre)

It is intended as a starting point, rather than a ready-to-use method

Avenues for de-idealisation

Assumptions made before

1. Model outputs are **point-valued**
2. There is a **best model** to use as a centre
3. We should include models in order of their **Euclidean distance from the centre**
4. We should form **intervals** from model outputs

Avenues for de-idealisation

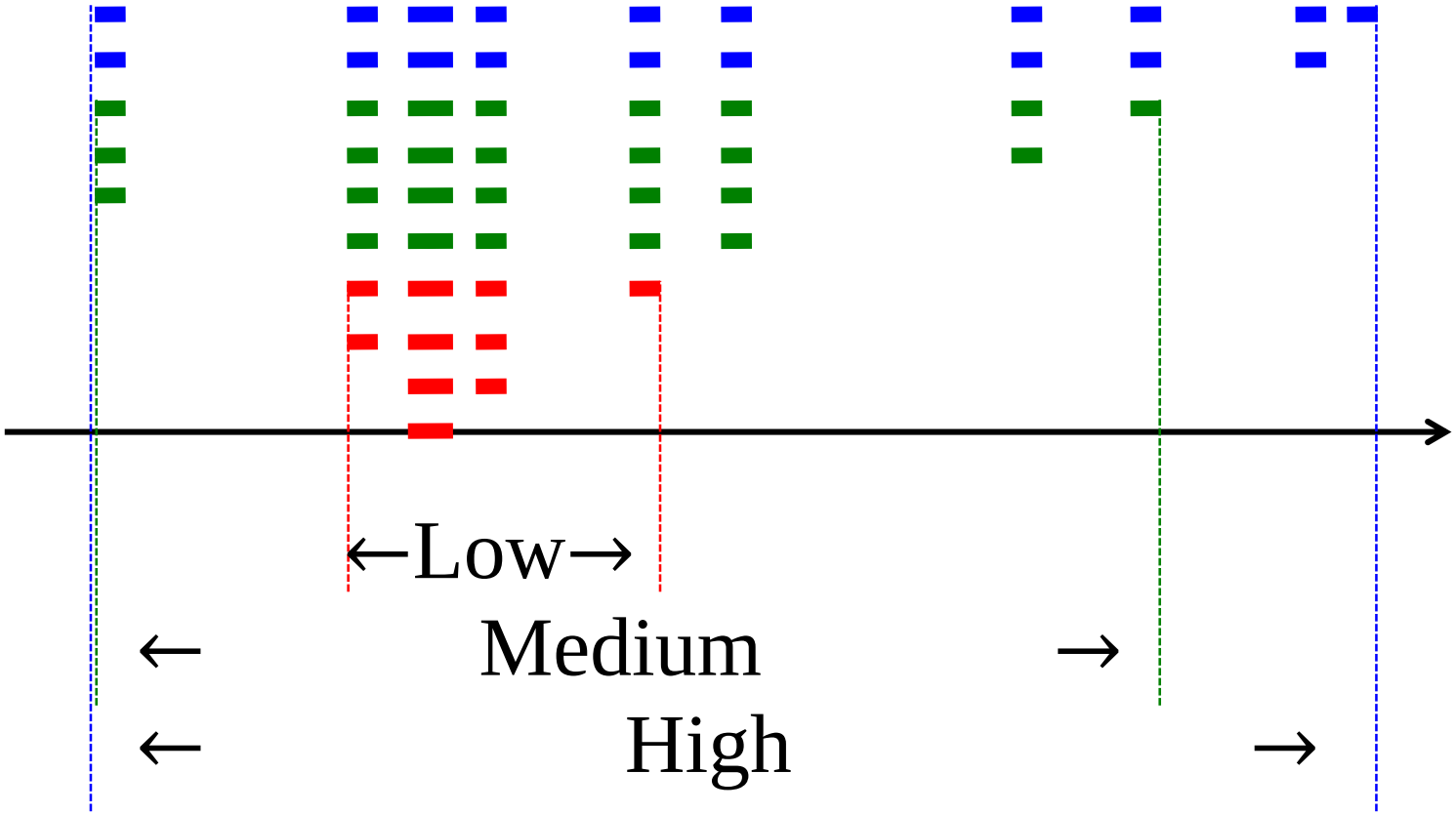
1. Model outputs are **point-valued**
   - We can run the same method, but with error bands, or even distributions, on each model output
2. There is a **best model** to use as a centre
   - We could centre on an interval, or set, independently of the above w.r.t. other points

Avenues for de-idealisation

3. We should include models in order of their **Euclidean distance from the centre**
    - If we can reliably **order** the models, we can include them in order
    - This is *more difficult* than simply finding the best, so not more general
    - It is still *less demanding than averaging*, however, which requires cardinal scoring rather than merely ordinal ranking

Avenues for de-idealisation

4. We should form **intervals** from model outputs



Alternative: statistical methods

- Fit a distribution over model outputs, use confidence intervals for “confidence levels”
  - IPCC 2013
- Surely this is obvious?! Why so late…
  - Which distribution? No good reason here to assume Gaussian
  - Do we assume models of equal value?
  - Naïve equal-weighting is sensitive to the number of models

Alternative: non-discountable envelope

- In the worst cases:
  - No best model
  - No scoring rule can order models
  - Do not trust statistical analysis

- Stainforth, Allen, et al. (2007) argue that this is the case for climate models used by the IPCC (in the CMIP5 ensemble)
  - Use *only* the outer envelope of all model results
  - Regard it as an estimate of the maximum uncertainty range (though likely a low estimate!)
  - Treat everything within as “non-discountable”

4. Open questions

- How to build in rich information from model validation and comparison studies?

- How do we connect more directly with literature on adequacy for purpose?

Thanks

References

Bradley, Richard. 2017. *Decision Theory with a Human Face*. Cambridge University Press. https://doi.org/10.1017/9780511760105

Bradley, Richard, Casey Helgeson, and Brian Hill. 2017. “Climate Change Assessments : Confidence, Probability, and Decision.” *Philosophy of Science* 84 (3): 500–522. https://doi.org/10.1086/692145.

Frigg, Roman, Seamus Bradley, Hailiang Du, and Leonard A. Smith. 2014. “Laplace’s Demon and the Adventures of His Apprentices.” *Philosophy of Science* 81 (1): 31–59. https://doi.org/10.1086/674416.

Frigg, Roman, Leonard A. Smith, and David A. Stainforth. 2015. “An Assessment of the Foundational Assumptions in High-Resolution Climate Projections: The Case of UKCP09.” *Synthese* 192 (12): 3979–4008. https://doi.org/10.1007/s11229-015-0739-8.

Hill, Brian. 2013. “Confidence and Decision.” *Games and Economic Behavior* 82: 675–92. https://doi.org/10.1016/j.geb.2013.09.009.

Stainforth, David A., M.R. Allen, E.R. Tredger, and Leonard Smith. 2007. “Confidence, Uncertainty and Decision-Support Relevance in Climate Predictions.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365 (June): 2145–61. https://doi.org/10.1098/rsta.2007.2074.

Thompson, Erica, Roman Frigg, and Casey Helgeson. 2016. “Expert Judgment for Climate Change Adaptation.” *Philosophy of Science* 83 (5): 1110–21. https://doi.org/10.1086/687942.

Wüthrich, Nicolas. “Conceptualizing Uncertainty: An Assessment of the Uncertainty Framework of the Intergovernmental Panel on Climate Change.” in *EPSA15 Selected Papers: The 5th conference of the European Philosophy of Science Association in Düsseldorf*. Springer, 2016. https://doi.org/10.1007/978-3-319-53730-6_9

A draft of the accompanying paper to this presentation is also available upon request.