

# Expert 1 – Fry

P1 · 40,000 nightly tests, acting on any result at  $p < 0.05$

## Model answers

### 1. "P1 WILL RAISE FALSE ALARMS BECAUSE..."

...a five per cent threshold accepts a one-in-twenty chance of finding evidence for something that is not there. Run 40,000 tests a night and roughly 2,000 will clear the bar by chance alone. The system then acts on them within 24 hours, so the false positives become policy before anyone can check them.

### 2. "BEFORE ACTING ON A FLAGGED RESULT, FRY WOULD REQUIRE..."

...a hypothesis stated in advance rather than one found in the data, and independent replication. Fry's examples are the point: Peto's aspirin trial "failed" for Libras and halved risk for Capricorns when he split patients by star sign, and power posing collapsed when others tried to repeat it.

### 3. "THE PILOT FINDING IS WEAK GROUNDS FOR ACTION BECAUSE..."

... $p = 0.00001$  and 0.3% are answers to two different questions. The p-value says the effect is very unlikely to be zero; the 0.3% says it is tiny. With 38 million records almost any difference clears significance, so strength of evidence has been mistaken for size of effect.

#### LISTEN FOR

"One in twenty." Multiplication of tests. The distinction between a significant result and a large one.

#### COMMON MISCONCEPTION

That a very small p-value means a very important finding, or that more data fixes the problem. More data raises the number of tests and shrinks the effect needed to pass – it makes P1 worse, not better.

## Expert 2 – Kahneman

P2 · 3,141 counties ranked monthly, best and worst 50 funded automatically

### Model answers

#### 1. "THE COUNTIES AT BOTH EXTREMES OF THE RANKING WILL MOSTLY BE..."

...the smallest ones – the same rural counties appearing in the best 50 and the worst 50, often in consecutive months. This is exactly Kahneman's kidney-cancer map, where the highest and the lowest rates both sat in sparsely populated rural counties.

#### 2. "THAT HAPPENS BECAUSE..."

...small samples produce extreme results more often than large ones. A county of 2,000 people swings wildly on a handful of cases; a county of two million cannot. The ranking is therefore measuring population size as much as health, and the extremes are an artifact of the method rather than a finding about the world.

#### 3. "PUBLISHING AN EXPLANATION FOR EACH RESULT IS DANGEROUS BECAUSE..."

...a random event does not lend itself to explanation, but the mind supplies a cause anyway. Once an official story is printed – clean rural living, or poor rural access – the noise becomes an accepted fact, and funding follows a pattern that will not repeat next month. Kahneman's rule: no hammer, no causation.

#### LISTEN FOR

The word artifact. Both extremes, not just one. The move from "small sample" to "so the explanation is invented."

#### COMMON MISCONCEPTION

That rural counties are genuinely unhealthier, or that the data is wrong. The data is accurate; the ranking built from it is what misleads. Also common: treating this as bias in the researcher when Kahneman's point is the statistics themselves.

# Expert 3 – Luyendijk

P3 · The advisory board closes; one national index of GDP and inflation

## Model answers

### 1. "CLOSING THE ADVISORY BOARD REMOVES..."

...the only part of the system that could go and look: interviews, minutes, site visits. That is qualitative fieldwork, and it is how Karen Ho found what a trading floor is actually like – zero job security, a five-minute horizon – by sitting on one for years. No data stream reports a culture.

### 2. "THAT KIND OF KNOWLEDGE MATTERED IN 2008 BECAUSE..."

...the models assumed bankers would protect their own institutions. Greenspan later admitted he had "presumed" exactly that. Long-Term Capital Management was run by Nobel laureates and still collapsed. The mathematics conferred the authority of physics without the predictive power of it – Luyendijk's hubris.

### 3. "CALLING THE NATIONAL INDEX NEUTRAL HIDES..."

...a political decision about what counts. Weighting GDP growth and headline inflation while excluding household and care work, and giving housing prices no extra weight, decides in advance whose interests the policy serves. It is politics disguised as technocracy.

#### LISTEN FOR

Hubris and humility. "What the index leaves out is a choice." Fieldwork as a method, not as an opinion.

#### COMMON MISCONCEPTION

That Luyendijk thinks economics is useless or that numbers should be abandoned. His target is the prestige – the claim to exact-science status – and the fieldwork that prestige lets economists skip.

If the panel approves with conditions, the condition answering this flag is keeping a human board with fieldwork powers, and publishing what the index excludes.

## Expert 4 – Nature editorial

P4 · Licensed traces for 38 million citizens; predictions pushed back as alerts

### Model answers

#### 1. "THE PEOPLE MISSING FROM THIS DATA ARE..."

...anyone who leaves fewer digital traces: poorer households, the elderly, people paying cash, people without a smartphone or a data plan. Traces are left disproportionately by wealthier people in developed countries, so a national conclusion is biased before any analysis begins.

#### 2. "ONCE CITIZENS RECEIVE THE ALERTS, THE SYSTEM IS REALLY MEASURING..."

...its own influence. Algorithms change the behavior they observe, so what comes back is a response to the alert rather than organic social action. PredictaGov reads its own feedback loop and calls the result a finding about the population.

#### 3. "CONSENT AND DATA OWNERSHIP ARE A PROBLEM HERE BECAUSE..."

...these records were generated for commerce, not research, and there is no separate consent process. Three private firms hold the data, de-identification can be undone, and the citizens being governed by it never agreed to that use.

#### LISTEN FOR

Two distinct problems, not one: who is missing, and the loop. Data collected for another purpose. Validating measures against theory.

#### COMMON MISCONCEPTION

Reducing the whole flag to privacy. Privacy is one strand; the editorial's sharper claims are sampling bias and measurement validity – the traces may not measure what the model says they measure at all.

If the panel approves with conditions, the condition answering this flag is validating the traces against survey or interview data, and suspending the citizen alerts.