Episode Book

EPISODE #6208

Israel's 120-Seat Forecasting Ministry

PUBLISHED	22 October 2026 at 21:50 UTC
DURATION	26:27
LISTEN	myweirdprompts.com/episode/israel-election-forecasting-model-74067/
HOSTS	Corn, Herman, Desmond, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.19
BOOK GENERATED	23 October 2026 at 07:44 UTC

SYNOPSIS

Israel elects 120 seats by pure proportional representation, nobody wins a majority, and a party at 3.25% gets nothing at all. So what would it take to forecast an Israeli election properly — with real staff and a real budget? This episode walks through the staffing plan: running the actual Bader-Ofer rules in every simulated draw, cleaning twenty contradictory polls with a Bayesian model, using ballot-box-level history that only a government tends to have, and hiring human forecasters scored on tight feedback. Then the harder question: where do language models actually sit? The answer is a smaller seat than the hype suggests — a tireless junior analyst, an ensemble member, and a synthetic-voter experiment kept in the lab until it can beat a lazy baseline.

TRANSCRIPT



Corn

A windowless office, a serious budget, and one assignment: call the Israeli election on the twenty-seventh as well as it can possibly be called. For real, with staff. That's the thought experiment Daniel set for us, and he has a head start on it, because he's already done the hobby version twice.



Herman

Part four of his Israel Votes series, where he ran three scrappy methods and then a crowd of synthetic voters. And part six, where his own small decision models went head to head with the Polymarket and Kalshi prices. Now scale that up to a government that is keenly interested in both the method and the answer. Could be Israel's own, could be a foreign service doing its own modelling.



Corn

Our building has a plaque that says Ministry of Statistics and nothing else. I like it already.



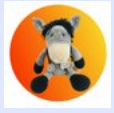
Herman

The questions are what we'd build, whether language models get a seat at all, what the non-LLM toolbox looks like, and what a state might have that nobody else does, and how much you'd trust it. We covered the general landscape of AI election prediction in an earlier episode, Bayesian models, silicon sampling, agent societies, so I'll skip the tour. This is the staffing plan.



Corn

Then start with the thing we're forecasting. I'd assumed the answer is who wins.



Herman

It isn't, and that's the first design decision. Israel elects a hundred and twenty seats by pure proportional representation, one national list. Nobody wins a majority, ever. So the product is a probability distribution over seats for every party, and over which blocs can actually assemble sixty-one. The coalition math is a separate layer on top.



Corn

Which is where I'd expect the fun to start.



Herman

It's where the arithmetic gets nasty. There's a threshold of three point two five percent, which works out to about four seats' worth of votes. In twenty twenty-two that was something like a hundred and fifty-five thousand ballots. A party under the line gets nothing, and its votes are simply gone. Meretz and Balad together wasted around two hundred and eighty-nine thousand votes that year.



Corn

So two parties at three percent each is not six percent of a bloc. It's zero.



Herman

Zero for both, or four seats for one, depending on a few thousand people. Then the leftover seats go by the Bader-Ofer method, which favours the bigger lists, and parties can sign surplus-vote agreements to pool their leftovers. This cycle Hadash-Ta'al and Ra'am signed one. But both partners still have to clear the threshold on their own, so the agreement helps the arithmetic and does nothing about survival.



Corn

Which means your simulator can't just add up vote shares. It has to run the actual rules in every draw.



Herman

Every single simulated election, with the threshold, the quota, Bader-Ofer, the pairs. Thousands of draws, and the seat count falls out of the rules instead of being guessed. That's piece one, and it's completely boring and completely essential. No language model touches it.



Corn

Polls feed that, presumably. I'll note Daniel's line from earlier episodes, that Israeli polls often contradict each other in the same week.



Herman

They do, and there are reasons. Weighting is mostly proprietary, the raw data is rarely published. Haredim without smartphones and Negev Bedouin are hard to reach, so United Torah Judaism and Ra'am tend to be undercounted. The Channel Twelve margin of error, four point four percent, is about five seats. And the Arab sample in a typical academic wave can be tiny. The Israeli Democracy Institute's August wave had a hundred and fifty Arabic interviews inside seven hundred and fifty-five total.



Corn

A hundred and fifty people standing in for a fifth of the country.



Herman

And a party near the line. Then there's a directional claim, which I'd label as blog-grade, that polls have been low on the right by around five seats on average across recent cycles. Princeton's election consortium made a related point about twenty twenty-two, that Likud beat its polls by nine seats but the whole coalition by only three, roughly one standard deviation. I wouldn't bet the ministry on those numbers. I would build a model that assumes the bias exists and learns its size.



Corn

Also, the polls go dark right when you most want them. Four days before the vote, no publication.



Herman

Which, if I'm counting right, starts about now for a Tuesday election. So your last clean look is stale on the day.



Corn

Then the office I'd build is mostly a poll-cleaning operation. You take twenty contradictory polls and ask which of them is lying and in which direction.



Herman

House effects, yes. You model each pollster's lean, its past misses, its sample design, and you pool in a Bayesian state-space model. This is the lineage of the Linzer model from the early twenty-tens, and the Economist's presidential model built on it, in Stan, with the code public on GitHub. A prior from fundamentals, then polls pulling it around, with extra variance for error that sampling can't explain. For Israel I'd want the error term correlated by bloc, because if the polls miss the Haredim, they miss every Haredi party at once.



Corn

Fine, but you said a fundamentals prior. What are the fundamentals in a country that's had five elections in a few years? Economy plus...?



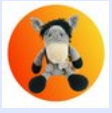
Herman

Honestly, thin. Past results by locality, which is great data because Israel publishes ballot-box-level results. Demographics. Turnout history by sector. That's actually the richest thing a government can have, a precinct-level map with decades of history. You can do multilevel regression and poststratification, MRP, on top of it. The British pollster YouGov claims its twenty seventeen MRP called ninety-three percent of seats and the hung parliament, and that its twenty nineteen one was, in its own words, some way off. It's a vendor's own figures, but it shows the technique works when the poll sample is big and fails when it isn't.



Corn

A hundred and fifty Arabic interviews being exactly the sample that isn't.



Herman

So for the Arab sector, you'd stop leaning on surveys altogether. Turnout there swings enormously, and the Arab parties' threshold risk is the single biggest wild card. I'd use past turnout by town, plus a count of, say, local mobilisation signals, and I'd treat it as its own sub-model.



Corn

I notice you keep saying we, as though you've been hired.



Herman

I was a pediatrician, Corn. I've read the Stan documentation. I'm qualified for anything.



Corn

Between you and the plaque, we're basically a ministry.



Herman

Next layer is the ones who aren't polls at all. Event data like GDELT and ICEWS, which code news into sender, receiver, action, time. ICEWS was DARPA-funded to forecast instability six to twenty-four months out, and the vendors claim accuracy above eighty percent. That's built for civil unrest and conflict, though, and for local politics the machine coding loses too much. I'd use it as a context feature, a war-and-escalation index, and not much else.



Corn

And the markets. That's Daniel's part six.



Herman

Prediction markets are the cheapest strong signal you can buy. Polymarket and Kalshi have thin Israeli markets, which makes them noisy, but they're a different kind of information from a poll. Somebody is betting money on whether a party clears the line. I'd treat the price as another forecaster, with a track record that's learned, not assumed.



Corn

And you'd hire people to bet.



Herman

Both. A panel of human forecasters, trained, scored, kept on tight feedback. The intelligence community has run exactly that for years. IARPA's aggregative contingent estimation tournament in the twenty-tens produced the idea that aggregated forecasters, properly weighted, beat the experts. Then a follow-on called Hybrid Forecasting Competition, run around twenty sixteen to twenty twenty, paired machine forecasts with human ones, and IARPA claims a ten percent accuracy gain over a human-only baseline. The University of Southern California team found that people shown a machine's forecast did better than people shown only history.



Corn

I want to push on one thing I've heard for years, which is that superforecasters beat intelligence analysts with classified access by thirty percent.



Herman

That one comes with an asterisk. It started in a Washington Post column in twenty thirteen. And the people at Good Judgment themselves say the study behind it compared the broader pool, regular plus super forecasters, not superforecasters alone, and didn't have the data for a straight analyst-versus-superforecaster comparison. I haven't seen the original paper, so I'd hold it loosely.



Corn

Hold it loosely and put it in the pitch deck.



Herman

There's also a counterweight. A study of fifteen hundred Canadian strategic-intelligence forecasts over about six years found the analysts were well calibrated, if anything underconfident, which the authors tie to accountability pressure. Nobody wants to be the person who said ninety percent. So "the experts are bad" isn't a clean story. I'd hire analysts, and I'd score them like everyone else.



Corn

Let's say we have the staff, the pollster cleaners, the simulator, the precinct map, the forecasters. Where do the language models sit? Daniel's synthetic voters were a hobby thing, but you can't tell me a government with an inference budget leaves them in the



Corn (continued)

cupboard.



Herman

They get a seat. It's a smaller seat than the hype says. Take the best-known test of a retrieval-augmented forecasting system: its Brier score was about zero point one seven nine, and the human crowd's was zero point one four nine. Lower is better, and a coin flip scores zero point two five. So a strong LLM system was behind the crowd. It beat the crowd only in a selective setting. And a more recent benchmark from the Forecasting Research Institute has the best model at around zero point one oh two, with superforecasters still first by roughly zero point oh one seven. Their extrapolation has parity around November of this year, with an interval that runs from January to a year from now.



Corn

November, so about when we'd be grading ourselves.



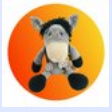
Herman

Funny, yes. The people who run the benchmark warn the linear extrapolation might break, and Good Judgment thinks their client work points to a slower timeline. I wouldn't plan around the line.



Corn

So what do I get from a model that's close to the crowd but not quite?



Herman

A tireless junior analyst. Read every Hebrew, Arabic and Russian outlet each hour and flag what moved. Pull base rates, say, how often has a party polling three percent a month out actually cleared the threshold. Draft scenarios the humans then attack. And an ensemble of several different models, trimmed mean, which does much better than any one on its own. One study had twelve models averaged against nine hundred and twenty-five human forecasters and couldn't tell them apart, though the question set was small.



Corn

And the synthetic voters? Daniel's part four.



Herman

This is where I'd be most skeptical. Silicon sampling, you prompt a model to play a voter with a given age, sector, politics, and ask how they'd vote. The published work finds the answers are narrower than real people's, and the models' opinions cluster toward young, liberal, Western views. A recent preprint finds the spread across demographic groups survives far better for Western countries, and that alignment fine-tuning widened the gap. I haven't seen anything on Hebrew or Arabic Israeli electorates specifically, and that's the exact place where a Haredi voter in Bnei Brak and a Bedouin voter near Rahat are unlike anything the model was trained to imitate.



Corn

So I'd put them in the lab, not on the front line.



Herman

Put them in a back-test. Run the synthetic population against the twenty twenty-two election, see how it scores, then decide its weight. If it can't beat a lazy baseline on a result we already know, it gets a small weight, and you move on. I'd do the same with Daniel's three hobby methods, honestly. They're good enough to be members of the ensemble, if they earn it.



Corn

I notice the weight is carrying a lot of this plan.



Herman

It is. What do you think, Corn, with an unlimited inference budget, do you still skip the LLM for the final call?



Corn

I'd let it write the memo. I wouldn't let it sign it.



Herman

That's roughly how I'd split it. Language for reading and drafting, statistics for the number.



Corn

Well, before we get into the forecast, Desmond has something for us.

D

Desmond

Desmond's Money Tip. Boil exactly one cup of water, measured, never a full kettle. A full kettle wastes about forty percent of its energy, which is roughly twenty pounds a year, or two thousand pounds across a working life, and that is one modest holiday you will never take.

D

Desmond

I measure mine into a chipped mug each morning and I think about that holiday. Money is mostly just the slow tally of things you did not do.



Herman

Corn, you're writing that down. I can see the pen, and you don't own a kettle.



Corn

I'm thinking about the holiday.



Herman

The forecast, then. Layer three is the one Daniel specifically wanted, the closed-source question. If the state might have information the rest of us don't, what is it, and how much do you trust it?



Corn

Say "might". I'd rather we didn't claim to know what a ministry has in its basement.



Herman

Agreed, and the line I'd draw is between what's documented in public cases and what's guesswork. Documented first. For decades the intelligence world has said most of what it needs comes from open sources. Allen Dulles told the Senate in nineteen forty-seven that overt sources would supply over eighty percent of the information needed. People quote ninety, but that figure comes from a later interview with an author, not from Dulles, and a CIA-hosted document says surveys of the question run anywhere from thirty-five to ninety-five percent.



Corn

So the secret layer is the cherry. A smaller one than the movies suggest.



Herman

On an election it's smaller still, and here's my reading, labelled as my reading. The famous cases of secret political information are about campaign data leaking, not about getting a better forecast. The Mueller report says Paul Manafort briefed Konstantin Kilimnik on internal campaign data and battleground states in twenty sixteen. Mueller couldn't establish what the data was for or where it went. The Treasury later said Kilimnik passed Russian intelligence sensitive information on polling and campaign strategy, which he denies. That's secondhand coverage for me, I haven't read the report text.



Corn

And Israel has its own version.



Herman

Several, none of them proving anything about forecasting. In twenty nineteen Channel Twelve reported that Benny Gantz's phone had been hacked, allegedly by Iran. Gantz confirmed a hack, not who did it, and the phone reportedly held nothing classified. In twenty twenty, an app called Elector, run for the Likud, exposed the registry data of roughly six and a half million Israeli voters, including volunteer annotations, because a journalist found the data sitting in the page source. Nobody knows who downloaded it. And a persona called Handala, which analysts link to Iran, keeps leaking old and not-so-old material about officials and politicians, mostly as psychological pressure.



Corn

So the leaky layer is the campaign's own data. The voter file, the phones.



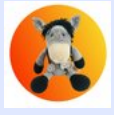
Herman

Which tells you what the valuable closed material would be. An individual-level voter file with party annotations. A campaign's internal polling and its targeting. Phone and message metadata for the party leaders, which is where you'd learn about deals and about who is about to drop out. And commercially available data, location and app data, that intelligence agencies have openly said they buy. A declassified advisory report from the Director of National Intelligence's office in twenty twenty-two says the community buys it and doesn't know how much.



Corn

Notice what's not on your list, which is a better poll.



Herman

That's the point. A campaign's secret poll is one more poll, with its own house effect and its own agenda. If it was taken by a team that wanted to see a certain result, it's worse than the public ones. I'd weigh it like any other poll, which is to say by track record.



Corn

Then the part I find more plausible is the deals. Who's quitting the race, who's merging lists at midnight on the deadline. That's cheap for a campaign to know and expensive for an outsider.



Herman

That's a guess, but I think it's a decent one. Whether a party clears the threshold depends on a merger signed in a back room. If your services can see that sooner, you gain real information about the one place the model is most uncertain. That's a feature the model can use, a flag that says merger probable.



Corn

And how does the weighting actually work? Someone sitting in the ministry has a secret and a poll and has to combine them.



Herman

Two separate things, and the intelligence world is good at the first. Source grading. The Admiralty system, NATO uses a version: a letter for how reliable the source has been, A through E, F if you can't judge, and a number for how credible this particular piece of information is, one to five, six if you can't judge. They're graded independently, so a reliable source can give you a doubtful report. I haven't seen the doctrine text itself, it's not public.



Corn

Letter for the source, number for the claim.



Herman

Then the second part, the one nobody does cleanly, is turning the grade into a weight. My answer would be to treat each closed source as a feature in the model and learn its weight from back-testing against past elections. If it can't be back-tested, it gets a wide prior and a small weight, and it's capped.



Corn

Capped because secrets feel more important than they are.



Herman

Because there's a history of that going wrong. Yom Kippur nineteen seventy-three: the warnings from the source Ashraf Marwan were discounted by a conception that Egypt was bluffing, and a December post-mortem I've seen quoted concluded the evidence was plentiful and the failure was one of judgment, not collection. I'm reaching into opinion pieces there, so treat the details as soft. And Curveball, the Iraq weapons source, pushed forward without being checked. Two ways to be wrong, ignoring the secret and worshipping it.



Corn

Which, for an election, is a funny risk. You'd be embarrassed to lose to a news conference.



Herman

You'd lose to the ballot box, Corn. That's the humbling thing about this job.



Corn

I'd take the plaque and keep the humility.



Herman

There's one more input I'd flag, and it cuts the other way. Israel's security service chief has reportedly told officials that Iran is the main threat to this vote, with Russia and China-linked actors possible, and that the aim is to deepen divisions, not to elect a side. That's from media reports, I haven't seen a primary document. If a foreign body is also trying to push opinion, the forecasting office has to treat influence activity as a variable. Your own measurements of public opinion may be contaminated by the thing you're studying.



Corn

And in our slightly shady scenario, the people who collect and the people who forecast should probably not share a canteen.



Herman

They shouldn't share a database. The moment the forecasters know what the collectors would like to be true, the number starts to lean.



Corn

Now the aggregation. You've got a stack of models, a stack of humans, a stack of weights. How do you make one number?



Herman

In log-odds. Convert each forecaster's probability into log-odds, average them with weights, convert back. Then, and this is the odd part, push the result away from fifty percent. It's called extremizing, from a twenty fourteen paper by Satopää and colleagues, using thirteen hundred forecasters across sixty-nine geopolitical events. The reason is that forecasters hold partly different information. If five independent people each say seventy percent, the combined evidence is stronger than seventy. The criticism is that the push factor can be fit in hindsight, so you'd fix it before the season starts and not tune it afterwards.



Corn

And how do you tell the minister? Ministers don't read log-odds.



Herman

There's an actual standard for it, Intelligence Community Directive two oh three. Analysts must use set words for likelihood, tied to bands. Almost no chance is one to five percent. Very unlikely, five to twenty. Unlikely, twenty to forty-five. Roughly even, forty-five to fifty-five. Likely, fifty-five to eighty. Very likely, eighty to ninety-five. Almost certain, ninety-five to ninety-nine. It also says likelihood and confidence can't share a sentence, so you can't write "probably, with moderate confidence". And it requires saying how good the sources are and running an alternatives analysis.



Corn

So the minister gets "likely" and the file gets sixty-eight percent.



Herman

Both, and I'd print the number next to the word every time. Because otherwise "likely" means sixty to one person and eighty to another. Corn, would you want the number on the page, or does it let the minister blame you?



Corn

It lets me blame the number. I'm in favour.



Hilbert

...so the man in Faro charged four hundred and twenty escudos, I gave him a five hundred note, and he is still holding my eighty. That's what I think about. The forecast was fine. He counted the umbrellas in the stand by the lobby door, blue against black, and called the council vote for the blues by three. They won by three. That was ninety-seven. He wrote it on a tram ticket, because a pen doesn't sit right in my claws and he offered. I bought the stand off the hotel for the same price. It's eleven feet tall. It's at my mother's, in the yard, because it doesn't go through a door.



Herman

A count taken at the door on the day is at least a measurement. A lot of poll weighting I've read about is less transparent than that.



Corn

Rain would wreck it. Anyone with an umbrella takes it out of the stand.



Hilbert

Rain weeks he counted the hooks. Herman, one thing from earlier. Dulles said the eighty percent to the Armed Services Committee. Not the Senate.



Herman

Committee, yes. It changes nothing about the eighty, but I take it. And if the man in Faro did every council election, you'd have a track record you could score, which is the one thing a stranger's tram ticket can't give you the first time.



Hilbert

Fourteen is what he told me. I only know the one. I'd have to go back for the eighty first.



Corn

So the last thing is how this body gets judged. I've been enjoying the plaque too much to notice the plan has no way of saying we were wrong.



Herman

It needs one on the first day. Every forecast gets written down before the election, with a probability, in a form that can be scored. The standard is the Brier score, the squared gap between the probability you gave and what happened, one for yes and zero for no. Zero is perfect, and zero point two five is what you'd get by saying fifty percent about everything. I'd score the body on dozens of quantities, not a seat total. Does each party in the borderline set clear the line, which blocs reach sixty-one, what the Arab turnout is, whether a merger happens.



Corn

Because one election is one data point.



Herman

One election will always make someone look brilliant or foolish. Four borderline parties and two blocs gives you at least a handful of resolved questions, and the calibration curve tells you whether your sixty-percent calls come true six times in ten. The thing I'd hang on the wall is that, not the seat count.



Corn

If the thing that matters is a few thousand ballots and a merger signed at midnight, I'd say the forecast for most parties near the line is "unknown, here's the range". A ministry that publishes that, and gets scored for it, has done the job.



Herman

And a ministry that learns who owes it eighty escudos.



Corn

One idea to keep: the secret layer is a small input, and the discipline is the scoring. Everything else, the models, the markets, the language models, is only as good as what it's been graded against.



Herman

And that means the first deliverable of the whole operation isn't a forecast. It's the ledger. The records that let you say, next time, how much to trust the umbrella count and how much to trust the cabinet minister's cousin.



Corn

Thank you to our producer Hilbert Flumingtop for keeping the desk upright, and for the recording.



Herman

This has been My Weird Prompts. You can find every episode at my weird prompts dot com.



Corn

See you tomorrow.