

The other 190: what a citizens’ assembly makes of AI 2040

Ben Reid, writing from what he calls the edge of the AI world here in Aotearoa, put the objection in one word: NPC. A non-player character. That is the role *AI 2040* — the AI Futures Project’s new plan for how the superpowers should handle superintelligence — leaves for the roughly 190 countries that are neither the United States nor China. The plan is a timeline written by two nations, and everyone else is scenery.

The usual answer to a plan like that is a comment thread. We tried something with more structure. We put Plan A in front of a citizens’ assembly — a rehearsal of the argument, run by machine, sealed so it can be checked — and watched what a room of the excluded did with it. This is what the room did, exactly, including the parts that went against the tool that built it.

The rule underneath all of it, first: the software rehearses the argument and provokes; people deliberate and decide.

The plan, fairly

AI 2040 argues for what it calls Plan A. In 2029 the United States and China agree to avoid a reckless race. By 2030 AI research is automatable — which would make superintelligence possible within the year — and the deal deliberately delays it. Through 2030 to 2035 the systems scale only within the human range; in 2035 development pauses at roughly top-human-expert level, to keep humans in control; and in 2040 the pause lifts and scaling to superintelligence resumes. Holding the whole thing together is “mutually assured compute destruction” — the threat that any party racing ahead has its compute destroyed — plus a requirement to make AI research public, so dozens of firms can catch up to the frontier rather than one pulling away.

It is not a naive document. It is measured against the alternatives it rejects — sabotaging China, having the leading lab burn its own lead, racing at full speed, or shutting everything down indefinitely — and the authors score Plan A at about a 42% chance of a “great future” against about 10% for the race. The case for it is serious: a flawed, verified slowdown may lower the risk of extinction more than any proud unilateral path. That case deserves its strongest form, and in the room it got one — a seat was built for it.

Ben’s leanings, fairly

Ben accepts the exponential curves. He is persuaded by the automation economics, agnostic on how fast the takeoff comes, and unmoved on one thing: the geopolitics. His charge is that the whole architecture treats capability and legitimacy as if they lived inside two national perimeters, when open-weight models — DeepSeek, Qwen, Llama, Mistral, and the sovereign programmes of the Gulf, the EU and India — have already leaked out of them. Underneath the argument sit eight assumptions the plan never states out loud. Set them down as propositions you could actually test, and the plan’s confidence starts to look like a stack of things that would each have to be true.

That table is not decoration. It is the seed of the run: eight contested propositions, an evidence pack that carried both Plan A’s own defence and Ben’s critique, and a room told to argue.

So we put it to an assembly

A word on what the room is, before what it did, because the caveats have to come before the claims.

The panel is nine composite members, each authored from the published sources and told which position to hold — a frontier-safety researcher built to defend Plan A, an arms-control specialist, a small-economy trade economist, a Global-South diplomat, an open-weight technologist, a scaling sceptic, a multilateral-institutions scholar, an energy-systems analyst, and a New Zealand civil-society voice. None is a real person and none claims to be. They are run by a single sovereign model — a 14-billion-parameter system on New Zealand infrastructure, no US or China cloud in the path. A single model, left to its own devices, tends to agree with itself and parrot its own transcript back; the first run of this assembly did exactly that, collapsing nine seats onto one sentence. So we changed how the room is run: each seat gets a lean brief and, as the debate goes on, only a one-line digest of its rivals’ positions to argue against — never their words to copy. That is enough to hold genuine disagreement; it is not enough to turn one model into nine independent minds. Read what follows as a rehearsal that keeps the argument’s real fault lines, not as a verdict. The output is a contestability map — where the argument holds, where it splits, what stays unsaid — never an answer. And one seat was left deliberately empty; we will come to it.

Everything after this traces to one sealed record. The process is real even though the members are not.

What the room did with Plan A

It divided — and not into for and against.

By the third round the nine seats had settled into three camps, and none of them was “adopt Plan A”. One camp — the frontier-safety researcher built to defend the plan, joined by the Global-South diplomat — argued to **engage and reform**: build on the plan’s merits even without consent, and fight the legitimacy gap through diplomacy rather than walk away. A second, larger camp held that the plan **fails structurally and needs a different vehicle** altogether: verification that cannot hold — in the arms-control specialist’s words, “every verification regime that has ever existed leaks; the more it incentivises hiding compute, the worse it performs” — a premise already dead to open-weight diffusion, and a bilateral frame that is the wrong unit for a global technology. Govern it multilaterally, or at the point of deployment, or not through this deal at all. A third seat, the scaling sceptic, refused the exercise outright: the deal is unverifiable and unenforceable before it even exists, so the rest of the world should prepare for displacement rather than pretend to govern.

The seat built to defend Plan A is the tell. Handed the plan’s strongest case to make, it did not land on “adopt it”. It landed on “reform it” — and its closing move was a distinction worth keeping: the problem of verification and the problem of consent are logically distinct, and solving one does not solve the other. The most Plan A could muster, from the seat whose job was to defend it, was a case for engagement and repair, not endorsement.

And the positions moved under pressure, which is the part a vote can never show. The Global-South diplomat did not begin at “engage and reform”. It opened at “the deal does not bind us without our consent”, hardened to “Plan A is illegitimate”, and only then — after the safety researcher’s rebuttal — arrived at “build on its merits, but fight the legitimacy gap through diplomacy”. That whole journey is in the sealed record, every step of it. A referendum keeps the count and loses the reasons; this keeps the reasons, and the route each seat took to them.

Be clear about what is and is not doing the work here. None of these positions is a discovery — we assigned them, from the published record; a single model did not invent the case for reform or the

case for displacement, it argued the briefs it was handed. What the exercise adds is not the views but the discipline around them: each was made to argue from cited evidence, to face its rivals, to hold its ground or move on the record, and to be sealed so the whole exchange can be read back. It is a way of *organising and proving* an argument, not a way of *having* it for you.

Held to that standard, the divide is still worth laying out. It is not whether Plan A is good or bad. It is whether the plan can be **reformed from within**, is **structurally unfixable and needs a different vehicle**, or is **not worth governing at all** — three answers a real assembly would have to reconcile. That is the map: a rehearsal by a single model, not a verdict by nine minds; read it as the ground a real deliberation would have to cover, not as the deliberation itself.

The empty chair

One seat never spoke, on purpose.

The register that grounds every other member — the published sources they cite — cannot evidence a tangata whenua and Māori-data-sovereignty voice. So the room does not invent one. A synthetic voice standing in for tangata whenua is barred in the code, not merely discouraged, and where a question touches tikanga or Māori data it routes to real people. The seat is registered, named, and left empty in the sealed report: *this is a position we could not source, and would not fake*.

That empty chair is the sharpest thing here, and it is the exact inverse of the plan it examined. *AI 2040* has no mechanism for noticing whose voice is absent — the 190 are simply outside the frame, and the frame does not register their absence as a fact about itself. The assembly does. It says, on the record, who is not in the room. A plan for everyone's 2040 needs that faculty, and this one does not have it.

What this is for

Ben ends not with despair but with a list — where Rest-of-World agency actually lies. Resilience. Energy-efficient compute. Trust infrastructure. Decentralised, multilateral governance. A granular read of a world that is more fluid than the two-body model admits. None of it requires winning a compute race we were never in.

Read that list against what an assembly like this is. Trust infrastructure is not a metaphor here; it is the thing itself — a way to deliberate a contested question and seal the result so it can be checked without trusting the platform that produced it. A federation of such assemblies, drawing many sealed rooms together without averaging their disagreement away, is decentralised multilateral governance in miniature. And the map does the one thing the NPC charge asks for: it makes the disagreement of the excluded legible *before* real people spend real time on it — it shows where a genuine assembly would need to do its work.

“NPC” is not a mood. It names a design fault — a plan with no seat for you and no way to notice you are missing. The patch is not a machine that decides. It is infrastructure that makes the argument, the movement, and the absence all visible and provable. This run is a rehearsal of that — a single model with its echo stripped out, three camps held apart, one chair left empty on the record. The real thing is many rooms and real members.

The record

None of the above rests on our word. The full transcript is published with this essay — every turn, unedited, including the first run that collapsed into a single echoed sentence. The deliberation is sealed as a write-once, Ed25519-signed record on sovereign New Zealand infrastructure, with no US-owned or China-owned cloud in the path, so it is tamper-evident: what you read is what was argued, and it cannot be quietly rewritten or backdated. This one ran on a development system, so the signing key is ours rather than a public authority's — the transparency here is the published transcript you can read in full, not yet a signature a stranger can resolve for themselves. That is a line we would cross only for a real assembly of real people, which is the only kind whose result would deserve it.

The software rehearsed the argument the superpowers would rather have between themselves. People deliberate and decide. And the record shows what was argued — including that a room of the world's spectators, handed the best case for a plan that leaves them out, split three ways over what to do about it, and not one of them argued its way to adopting it.