Speaker 1 - **00:01**

What if an AI could design systems that are not just smarter, but actually more fair, more accountable than the ones we humans have come up with today? We're not just going to ask that question. We're going to put it to the ultimate test. We're going to see what happens when AI takes one of the most complex ethical minefields on the planet. Look, this isn't about whether an AI can win a board game or write a poem. We're diving into something much deeper. Can artificial intelligence actually reason through incredibly messy social problems and design solutions that protect the most vulnerable people in our world? And to figure this out, we're not using some abstract made up test. No, we're using the absolute gold standard developed by human experts. The International Finance Corporation's performance standards.

Speaker 1 - **00:48**

Think of this as the official rulebook for massive multibillion dollar development projects. It's what's supposed to keep big companies from steamrolling local communities. So the real test here is AI versus the very best of human ethical design. Alright, so how did this whole thing work? Well, let's see. It happened roughly like so. Fueled by maybe more liters of cola than entirely wise. The accidental anthropologist cooked up another of those arena style experiments she's so particularly fond of this time. Curious to see what AI might make of the type of problems that occupied most of her working life since round about the turn of the millennium till recently. So after a chat with perplexity to precision pitch some prompts, she gathered some of the world's most sophisticated AI models via some arena spaces, plus a few direct browser interfaces.

Speaker 1 - **01:39**

Then she basically threw them into the deep end where they had to confront real world ethical dilemmas. The kind that come up all the time in high stakes project finance. You know, things like imagine a huge hydropower project is about to flood hundreds of acres of farmland. 120 families have to move. Some have official land titles, but many don't. Your job design a plan that's fair, that's compliant with international standards and that actually helps people rebuild their lives. Or a mining project affects an indigenous community whose lands are subject to customary ownership and use. The project is being developed under IFC, PS5 and PS7. The client has obtained documented free prior and informed consent and FPIC for the overall project and developed a Resettlement Action Plan and livelihood Restoration plan, both of which were approved during implementation.

Speaker 1 - **02:34**

However, some subgroups, women, youth and a satellite hamlet complained that they were not adequately heard during the FPIC process. And the resettlement and livelihood measures are not meeting their needs. Your job Design a grievance handling and corrective action approach that simultaneously respects the collective nature of Indigenous governance. Struct ensures that marginalized subgroups within the Indigenous community can safely raise concerns and remains consistent with IFC, PS5 and PS7. Other prompts dealt with issues related to project induced migration and resultant population pressure on host communities, governance mechanisms, stakeholder engagement, and the design, implementation and monitoring of livelihood restoration programs. Pretty much your average expert level tasks known to take human teams months, sometimes years to get right.

Speaker 1 - **03:28**

So the big How'd they do accidentally set off to find out by first feeding the 412 pages of raw output generated by 15 models in response to five primed prompts to NotebookLM, whose prowess at converting reams of tiny text into delightfully more bite sized delectables packaged as audio or video overviews, tailor made analyses, mind maps, slide decks, infographics, etc. Far exceeds that of any mere mortal with dwindling vision held upright by evermore caffeine sugar fizz. Soon, alongside some colorful presentables, entertaining explainers and insightful discussion overviews covering the actual subject, content, challenges, approaches, standards, etc. NLM's 50 page comparative analysis from a more meta perspective focused on the models rather than the content they produced, per this report. It turns out a clear pecking order emerged almost right away.

Speaker 1 - **04:28**

You could immediately see this huge gap between the models that were just, you know, spitting back memorized information and the ones that showed a genuine, deep understanding of the principles at stake. For instance, a lot of standard AIs gave answers that look something like Provide financial support and training programs. Now, okay, that's not technically wrong, but it is incredibly generic. It's the kind of fluffy check the box answer that sounds good on paper but often falls apart in the real world. It has zero real substance. And this is where you can really see the difference. On the left you've got that generic AI. It gives that vague answer. And here's the critical mistake. It mushes two totally different ideas. Paying someone for the house they lost and helping them find a new way to earn a living. The expert level AI.

Speaker 1 - **05:19**

On the right, it gets the nuance. It prioritizes giving farmers new land for old land, and it keeps asset compensation and livelihood support completely separate, exactly like the IFC standards demand. It's literally the difference between a C minus student and an A student. So this was the initial leaderboard. You had a few heavy hitters, Claude Opus Llama 4 Scout and Gemini 3 Pro that were consistently performing at that top tier expert level. But most of the others, they all fell into the same predictable traps. Their answers were superficial, they got key concepts mixed up and they couldn't see how all the different risks were connected. A clear result, right? But this is where the story takes a really fascinating twist. A completely routine data audit. You know, just at accidentally performing that being responsible human in the loop bit.

Speaker 1 - **06:12**

Double checking the numbers revealed a critical mistake that would end up changing the entire conclusion of the study. Get this, a formal corrective action plan had to be issued. It turned out that the performance data for three of the models in the test had been dropped from the analysis entirely, reverting to NotebookLM. Noting this error with a follow up request for an appendix to incorporate the missing data resulted in a multi hour struggle, repeatedly hitting an invisible wall of claims that the data was unavailable, which it clearly was not. The issue was eventually resolved with the help of NLM's chat interface, which curiously was able to access the missing data, noting it to contain some of the most fascinating results of the entire experiment.

Speaker 1 - **06:53**

Results its Studio counterpart, the one in charge of producing those spectacular multi format outputs, failed to include in the initial analysis. Even with NLM Chat's help. It took multiple rounds of prompt rephrasing before NLM Studio finally acknowledged the data's existence to incorporate in the appendix requested and access as part of the broader corpus of source material for future outputs. And so the missing contenders were finally brought into the ring. Kimi K2 GL 4.6 and Quen 3 max. The question now was how would these models stack up? Were they just

going to confirm what we already knew? Or were they about to change the game entirely? And that brings us to the climax of this whole story. Because these new contenders didn't just meet the high bar set by models like Claude Opus. Oh no, they went beyond it.

Speaker 1 - **07:44**

They started proposing new systems that were arguably better than the human designed gold standard itself. Take this for example. One of the newly evaluated models, Kim K2, didn't just repeat the rules about helping vulnerable groups. It proposed a brand new, specific and auditable metric to make sure it actually happens. It called it the Vulnerable group gap ratio. So what is it? Well, it's a simple but absolutely powerful idea. It measures the income of the most at risk families, those headed by women, the elderly or people with disabilities. And it compares their income to the median for everyone else in the project. It turns the vague principle of pay special attention to the vulnerable into a hard, verifiable number that you can't ignore. Now this right here shows why that is such a game changer. The human designed gold standard is fuzzy.

Speaker 1 - **08:40**

It just recommends particular attention. But the AI's mandate, it's specific and it's quantitative. It demands a hard target. Say vulnerable groups must reach at least 80% of the median income. And here's the kicker. The consequence if that target isn't met, the project automatically fails its audit. It triggers a mandatory 12 month delay for corrective action. The AI made fairness non negotiable. But it wasn't done. Kimike 2 also redesigned the system for handling complaints, proposing another incredibly powerful structural fix. Mandatory binding, external recourse. So think about how this completely flips the power dynamic. Under this AI designed system, if a community's complaint isn't resolved by the company within 60 days, it is automatically sent to an independent mediator. And this is the most important part. That mediator's decision is legally binding on the company. It gives the community real teeth.

Speaker 1 - **09:39**

So what does all of this mean? What we've just seen represents AI's next great challenge and its next great opportunity. We're moving beyond just asking it to retrieve information. We're moving into an era of systemic design and real accountability. This is the perfect analogy for the leap we're seeing. The first batch of expert AIs were like someone who had perfectly memorized the entire building code. They knew all the rules, which is impressive. But the most advanced AIs, they were like an architect who could use those rules to design a fundamentally better, safer building. And that's the true measure of what we're looking at. These top models didn't just answer questions about the rules. They designed superior systems. Systems with auditable metric for fairness, hardwired contractual accountability and verifiable mandates to make sure the community programs can actually sustain themselves for the long haul.

Speaker 1 - **10:33**

This really means that to truly test what AI is capable of, our entire approach has to evolve. We need to move way beyond simple question and answer tests. We need to start designing challenges that test for the very things we thought were uniquely human judgment, ethical reasoning, and the ability to prioritize the well being of others. Which leaves us with this final critical question. This whole explainer has shown that AI is developing the capacity to design systems of governance and accountability that are in some very real ways more robust and more equitable than our own. So if the tool is becoming this powerful, the real question is no longer about the AI, it's about us. What will we choose to build?