How Ricursive Intelligence’s Founders are Using AI to Shape The Future of Chip Design

36 MIN · YOUTUBE · HTTPS://YOUTUBE.COM/WATCH?V=55LT52EVARM
https://youtube.com/watch?v=55LT52eVArM

Right now we can't have much of codeesign between chips and uh models because of this asymmetric design cycle for chips because it takes so long uh so much it takes so much time to design chips. The the cycle is there's a mismatch between how fast we can create the next generation AI methods and how fast we can build the next generation chips. But if we can make [music] our chips much faster then we can enable this co-design and co-evolution of workloads applications and chips all together.

[music] [music] This week on training data, we explore the biggest bottleneck holding back AI compute and the design of chips themselves. Our guests are the founders of recursive intelligence, Anna Goldie and Aalia Mirhoseni, the team behind Google's pioneering alpha chip project that helped design four generations of TPUs. They're now applying AI to the entire chip design process, transforming the industry from fab to designless.

Anna and Aalia paint a picture of what a Cambrian explosion of custom silicon means. We'll explore the holy grail topic of recursive self-improvement, how AI design chips unlock radically creative new chip designs, and why democratizing chip design could accelerate the path to super intelligence. Enjoy the show. Thank you so much for joining us. We're so excited to celebrate the announcement of recursive intelligence and congratulations on embarking on this new adventure. Thank you. Thank you so much.

Um >> to kick off, um you've said that chip design is the compute bottleneck to advancing AI. Can you share and paint a picture of what's happening now? Um what are all the bottlenecks to chip design today? I mean I think when we were saying that we were kind of alluding to this motivation that we had originally with our moonshot which was this observation that you know neural networks these concepts that have been around for decades but the AI that came out of it wasn't that effective until we had like more powerful computer systems and chips and we thought that now we that we had very powerful AI systems we could use those to kind of tackle the bottlenecks in chip design and you know create basically more effective compute.

So if you see these like scaling laws basically the more compute you apply to training a model or inferring with a model like the more intelligence you get out and we're seeing that there's this mismatch like we're using say GPUs are like originally designed for graphics processing but we're somehow repurposing them for crypto and then for training neural networks models. I guess they're good at like large matrix multiplies, but like if we could more better co-optimize the models and the hardware, we could get more effective compute and then we could push ourselves out on that those effective scaling laws.

>> You were the co-creators of Alpha Chip at Google, um, which I think is using four successive generations now of TPUs. Maybe take us back to that project. How did it get started? Uh, what were the key results you drove? >> Yeah, so we started a project uh in 2018. Um and we had some earlier version of placement that placement work that we were interested in solving and that was not about chip placement. It was about compiler and ma basically mapping neural networks to chips. Um so we did that project we got great results we published and we were like what are the like highest impact projects that now we can kind of push this further and have real world impact and that's where the chip placement uh came in in in the picture and we started the project working very closely with the TPU team at Google because back then we didn't know much at all or at all about how chip design is uh how the process is. So we worked very closely with them uh from very early on and uh at the time it's interesting because at the time AI was like was still like very talked about but nothing like the scale was nowhere near what it is today. So we had to really work through things and showing with showing them data and like constantly like iterating over our approach and hearing them what they needed from us, listening to them. um iterate through the process and eventually uh we bent from a research concept on chip placement all the way to something that was actually used in product and we could tape out and all that. I remember one of our very first meals together um when I it hit me that you were so customer obsessed and had so much empathy for the customer and I asked you how and why and you said well I've been working with the TPU team as our customer internally for so many years. So I think that really kind of gave you that other perspective. Um what were the TPU team's early reactions um to what you had created? I mean they were like super skeptical. [laughter] For example, like Aaliy and I we were you were researchers. Like we had read a bunch of research papers and we're like okay there's this half perimeter wire length that's what academics report results on. So like we made this RL agent that could optimize half perimeter wire lengths for a placement. And then we were kind of excited about sharing that with them. And we showed these results and they were like actually kind of like angry at us like why are you showing us these results? Like we don't care about half perimeter wire length.

like we want like routed wire lengths, congestion like horizontal and vertical congestion, timing violations, power consumption, area like >> and so you know we we just kept listening to them and we were like okay so what is the thing that matters I think another part is just to make your customer feel like they're part of it too. >> Yes. >> So for example we worked with them to create the cost functions that they cared about approximations of those. So like Mustafa who was on the the TPU team developed these very fast like congestion cost function. Um and then we could optimize against the end density and then wire lengths as well and then we show results on those and then we run the commercial tool and we show that this actually correlates to good results on the metrics that they do care about.

>> Yeah. >> Um so I want to talk about floor planning and and I come from an EDA family. Uh you know and floor planning is it is the crown jewel of EDA. Can you say just for our listeners that don't come from the chip design industry what what exactly that is and how your technical methods help solve it? Uh so floor planning is a process where we place the components of a chip onto the silicon and uh the complexity of this process comes from the large size of the input problems like for a block of a chip not the entire chip like a single block in a chip the graph that we're optimizing has mill could have millions of nodes and all of these nodes need to be placed and routed and we need to make sure that all these constraints that earlier uh we were talking about like PPA power area performance are optimized and all the physical constraints are met. So the placement problem has to do with how we can do this such that all those constraints given by these like very small technology node sizes

given by TSMC and others are met. Uh so we're dealing with a combinatorial optimization problem. That's large scale but also all the metrics are hard to evaluate and measure.

>> Got it. And this traditionally was done together with EDA software like a cadence of a synopsis, right? How does your method differ from the classical methods? So what we de developed here was a learning based approach to the problem where we would train a reinforcement learning agent that would try out different ways to approach the placement problem and it would learn through interactions with this environment that we built to learn from the positive placement and also the negative examples and iteratively improve itself. Uh so one of the main differences between our approach and all prior approaches were uh this ability to learn from experience which would enable the model to self-improve just like a human expert that becomes better as as they solve as they solve more instances of the problem and harder problems. um our agent uh was able to kind of exhibit similar behavior and that was a very um major kind of uh delta and change between what we had and prior approaches.

>> Yeah. And I remember with I mean if you take Alph Go as an analogy move 37 was kind of shocking because it was just so different from how a human player would move. Are you seeing something similar happen with with floor planning and chip design? >> Yeah, we saw these very strange like curved placements. So they're donut shapes as well. I think humans would tend to make the macros very So macros are memory components that are larger and they make them very aligned and then they would put logic maybe in the middle and then wires would connect all of these components. But if you make the shapes curved, you can reduce the wire lanes which can reduce a power consumption and timing violations. But it's just the complexity of making this curved placement was like beyond what humans would have wanted to take on or it felt risky to them. So yeah, these aliens. [laughter] >> What was the moment you realized that this had the potential to transform the entire end to end process, not just the floor planning stage and when you decided that it had the potential to really be a company? There there were a few like milestones that uh first was like showing that it works that it gets the superhuman results on some blocks and then later on the chip was actually taped out and it came back and it was actually working which is a big a big uh kind of milestone uh because it can be really really costly if for some reason our AI missed something. Um so those were big moments. Uh the real impact and also on the algorithmic side when we saw this kind of self-improve improvement properties and the fact that models were our models were becoming better as they solve more problems like that that is like very cuz when we get to that point then there is no stopping for the AI.

they can see more data points and solve more instances of the chip optimization problem than any single human can ever do. Mhm. >> So that's that's the moment that you're like yes so [laughter] there's a lot more potential here but um alpha chip was targeting a module a a part a part of physical design uh but chip design as we know there there are many more to like stages to it and many more components and right now uh we think that everything from the state of AI and the capability of AI um both on LLM side and other like graph optimizations and other approaches and the way that we can really scale up these algorithms and enable synthetic data very large scale distributed computing everything is ready to tackle the end to end chip design optimization problem so that's why we're excited about doing the company right now super cool how do you think about getting the right training data for this market I think it's you know it is kind of the binding constraint to AI in many ways and is this a market where you can have you know enough synthetic data or

selfplay to be able to to bootstrap?

>> We're excited about synthetic data. So obviously there's some open source data but it's not that interesting or meaningful and you know customers are willing to share data with us but we don't want to train our models on that because we want to keep their data private and siloed from each other. But synthetic data is where we think there's the most promise. Like as Ali and I, we've been working on synthetic data approaches and for LLMs in various code domain tasks across like Claude and Gemini. And we see a big opportunity to like get a scale of data that would go far beyond what any customer could ever share with us, like orders and orders of magnitude more.

>> With the TPO program, um I guess at what moment did the chip experts stop doubting you? And as each successive generation came out, what was the progressive impact of what RL was able to drive in those designs? >> I mean, I think that our approach with the TPU team is like every week we would show them data over and over again >> and every week. >> Yeah. For every week for like a couple years. Yeah. Because I mean, this the stakes are very high here. Like if there's something wrong with this layout like the the setup on the TPU team is like this. There are human exp the the TPU is quite large and complex. So they divide it up into dozens of blocks and then each block is owned by a human or a team of humans and that's their responsibility to get this block right if anything goes wrong like it's their fault right and so they would generate their own layouts in collaboration with commercial tools and then we would show them our layout some AI generated layout it would look weird it would be curved and they would you know they would have to say okay like I'm picking this AI generated layout over my own layout u and I'm taking responsibility for that I think yeah requires a lot of trust. We have to be better in every single metric for them to choose that and we saw across each successive generation of TPU that we were being adopted in more and more of the chip block and more of the area and also that we were getting increasingly superhuman performance. I we published a alpha chip blog post I think September last year we showed this >> curves both say a word on that yeah on on the blog post in the superhuman performance. Yeah, I think like just that in across every single generation that we were used which I think there were three generations shown in the blog post but actually we're used in another generation after that was published every single one we're seeing more of the area and more like increasingly superhuman performance.

>> Yeah. So basically the delta between alpha chip layouts and the baseline layouts are also growing growing >> which is like it's a property of AI and how it scales with data. >> It makes sense right? We're alpha chip was trained on more and more >> TPU block so it gets better. >> Why is the company called recursive? >> Uh because we're recursive we're AI for chip design and chip design for AI. Uh so recursive self-improvement in terms of like the name the spelling of the name. Um so the name itself is recursive. So the initials R I recursive intelligence are the first two letters of the company's name.

>> What does recursive self-improvement like? Like like why why does that matter for you know the brother AI race like just say a word on I think this is this is a no problem. Just say a word on it. >> Yeah. So af chips are the fuel for AI and scaling laws are driving much of the progress in AI whether it's on pre-training post training test time and all that and so the faster we can make chips that are more custom or better designed for the AIs that we

run uh the faster we enable uh this efficient more efficient kind of compute and that kind bends the curve for our scaling laws. So that means we get to the next generation of AI uh faster and we can gen create design better AIs faster. And so this is a type of recursive self-improvement that we want to see because our AIs then can help our chips become better and we can design them faster and so on. So that is the recursive self self-improvement loop that we are going after. One of the visions that you have is to transform the industry from just fabless to designless. What does that mean?

>> So fabless was basically this concept of it used to be that people thought that no serious chipmaker could exist without their own fab. This was like obviously before like this multi- trillion dollar companies like Nvidia, you know, TSMC basically created this whole new world where we could have these incredibly valuable fabulous companies. So we think that there's an opportunity to create incredibly valuable companies that don't need to have their own in-house design teams. So many companies they serve these models at massive scale. They're spending like you know this I think a hundred billion dollar plus on AI inference alone like this year and it's growing rapidly. So companies would benefit from like maybe custom chips to serve their models or train them, but that requires enormous teams of like hundreds or thousands of human experts inhouse. Um, and we don't think that's necessary. So we want to move towards yeah designless paradigm. It's fascinating. We're seeing it come true with obviously Google and TPUs, Amazon with Tranium, also OpenAI and Broadcom and and maybe even Tesla. Do you think the future is then you know even today there's almost like model model application codees. Do you think there's going to be chip application codees and even individual companies will have multiple chip architectures as a result?

>> Yes, we think that there's there we are going to see more and more code design across the deep learning or AI stack from model and data to uh software and all the way to the chips. Um and we're going to enable that and code design is really the seek the secret or the path to more efficiency and more performance going forward. And uh but right now we can't have much of codeesign between chips and uh models because of this asymmetric uh kind of design cycle for chips because it takes so long uh so much it takes so much time to design chips. the the cycle is there's a mismatch between how fast we can create the next generation AI methods and how fast we can build the next generation chips.

>> Um but if we can make our chips much faster then we can enable this co-design and co-evolution of workloads applications and chips all together. One of the things that I loved that we talked about early on was that um the value you bring isn't just on reducing the cost um it takes to actually design a chip or the speed at which you can accelerate the chip design process though that is transformative in itself. It's actually to unlock completely new potential applications and maybe custom silicon as well alongside it. Um can you share a little bit about that? What is the Cambrian explosion that you might expect? I mean, we think computing is going to be increasingly ubiquitous in every aspect of our lives. So, like obviously there are things like ARVR, there's like maybe chips in space, even like hearing aids. Like there are experiences that aren't possible unless you can serve them at like sufficiently low inference like latency or at like low enough power. And we think that like custom silicon can enable these applications.

>> Yeah. They think that AI is going to be everywhere in every kind of experience, every every aspect of

industry and life going forward. And there are chips that would enable these AIs to run. And given the scale, we we would want them to run as efficient uh very efficiently and uh low power, highest speed, all of that. And custom silicon is really going to enable that. Um and we want to enable custom silicon for for basically any workload that is being run at sufficient scale.

>> Since we have the ear of whoever is listening on the podcast, who might some of your ideal customers be? Um there are a range of customers that we are uh envisioning. uh the the in the first phase of the company when we are f when we are building ways to accelerate dramatically accelerate the chip design process our customers would be chip designers um like Nvidia AMD ARM MediaTek and all all of these companies that are already designing their chips and all of them would want to faster design cycle and we can help them with with this uh technology Uh but we we don't want to stop there. Uh we uh we want to enable any customer that has a workload or a family of workloads that they want to run that that they're running it at sufficient scale and they would benefit from a from custom silicon. We want to enable them to have that without having having teams of hundreds to thousands of chip designers in their companies like so we I think that kind of goes back to this Cambrian explosion of chips that we can enable.

I'd love to talk about the kind of existing incumbents in chip design. Um so like I grew up in a family of of Cadence. Uh, mom and dad both worked at Cadence and I think chip design is a duopoly between Synopsis and Cadence. Um, and I think they're adding AI to their product suites. Um, how do you see your company playing out versus the incumbents adding AI? >> I think that we see ourselves as like sort of coming from the opposite direction like we're a frontier AI lab.

we come from like you know Google brain or like anthropic those kind of like backgrounds and we want to kind of rethink or reimagine how chip design can be done >> and we we think that fundamentally in order to be able to this isn't like a point solution thing where we want to replace some module one at a time with AI we want to like reimagine and co-optimize different stages and we think that an AI approach is necessary here >> what people from the chip design industry think of you guys is I would imagine there's a range of like from excitement to extreme skepticism.

Extreme excitement to extreme skepticism >> or extreme fear [laughter] fear. Um like yeah what do the shift design folks think of you all? >> I think a lot of them are excited to work with us as like potential customers and we're excited about them too and a lot of people are excited to come join us and work together >> but yeah of course like what we're doing is very ambitious so I'm sure many people are skeptical too.

>> Let's talk about that for a little bit. Like there was, you know, there's some internet uproar, you know, micro niche internet communities love getting spicy and there was like some spiciness around um alpha chip and like people from media saying like ah that's not it's not really real for xyz reasons. Like what what do you think was behind that and what do you think was valid in that criticism and what what do you think um is the important thing that they weren't realizing? Um so whenever AI goes to a new field and does something disruptive um we would see reactions like that and this is not necessarily just just for in our case it appears almost in every other >> the bitter lesson >> field it's the >> and we think the bitter lesson is part of it like we are true believers in the

bitter lesson and it's usually a a little challenging ing to kind of like um accept that a whole new technology or new way of looking into problems that people have been have worked on for decades is just coming and is solving things more end to end and um and these people like in this case Anna and I and the alpha chip team where we were coming from not from not an EDA background uh at least we had not worked on that's that's that problems space and now we could come come up with problems that were with uh solutions that were being productionized and all that. So, so there's always this kind of a reaction in the beginning but um somehow the true kind of impact of our work was was much bigger than just the problem that we solved. Um and I can mention some of those. For example, uh this bringing like doing reinforcement learning looking into these graph graph neural net optimizations uh were applied to many other problems across chip design including a best paper award in at DAC in 2023. Uh the first author of that paper is now is now our teammate at recursive and that was for physical design. We saw there its adoption in other stages of chip design like synthesis and so on. The work brought a lot of like attention to AI back in 2020 in in the chip design field and that was uh we think it's like one of the most impactful thing as a result of our project. Right now there are uh conferences like we we actually we are involved in one of the LLM AED design that are just just focused on AI approaches to chip design which back in the days was not not a thing. So, so we we are there's so many positive things happen as a result that we are grateful for that and >> I had a funny interaction at like ML for systems workshop which is this another conference that we had started in like 2018 and this person was like oh alpha chip inspired my entire PhD thesis and I was like oh that's wonderful and I was like so I'm just curious like how did you first come across the paper and he was like oh because of the controversy >> yeah [laughter] very amusing I think the thing that surprised me a bit was I thought that the people who would be upset by the work would be like physical designers whose jobs were at risk, right? And those and you know those physical designers were skeptical. It required a lot of data for them to be able to change their mind and accept that their method worked. But they actually weren't the people that were upset. It was people who had developed prior methods in the field.

>> And I think in retrospect that makes sense. There's something painful about like you pour like your like time, your like your human ingenuity, your like soul a bit into these methods and then people come from outside your field and then they're using sort of in some sense simpler approaches like they're using things at scale with data and compute and you're outperforming. It's like a little it's painful. >> Totally. >> But you know, everyone we can all you know build on top of this work.

>> Yeah. We can all adapt. Yeah. >> Can I ask you the opposite question? Um, so I love the bitter lessness answer. I I then have the opposite question, which is well the large language models became so good at coding for example, um, why won't they just naturally become good at chip design? Why does there need to be a specific chip design frontier lab? [snorts] >> Yes. So chip design has a a lot of components and some of them are um language and related to language and code but actually a big chunk of it is not related to language and code and it has to do with these uh different properties and graph structure of of the of the chip and all these constraints that appear like these are like really really hardcale combinatorial optimization ation problems that required a specific like the custom way of uh like specific kind of approaches from an AI perspective. Um and our approach here is to apply the right method to the to the uh to every problem and LLMs uh we love them so much and we're going to use them extensively. We are going to build uh LLMs that are really useful for some stages of the process but they're uh not going to be sufficient for all of this. Um so we are going to have our own AIS and specific optimizations for different stages and those are really important because uh unless we design

them those those uh kind of modules that are really fast really optimized we can't iterate around them with LLMs or with other methods.

So, so our approach is going to be a very hybrid uh kind of multifaceted uh AI that is going to after this uh but we are true believers in AGI and >> bitter lesson >> bitter lesson and we we just want to be on the frontier of that with better chips. >> What do you think a world with AGI looks like? It's so hard to imagine the world in 10 years or even five. But what what do you imagine for whenever that future is?

>> I guess AGI like what does that mean human level for everything? Like maybe you just have many employees that are just effectively you know a bunch of compute powering them. I think eventually like you know maybe we can all take a vacation but [laughter] >> it's just uh so it's going to be extreme like uh cases of productivity for every single human being. Um and as a result we can create a lot of uh economical value uh at a scale that is not possible today and hopefully distribute it at a scale that is not possible today. So uh I I have a very optimistic view or view about future about the future where we when we have AGI just lots of good things are going to happen and if you're careful about it we can distribute the value to to all humans and and that's great.

>> We're going to have data centers in space. >> Yes, why not? >> And is that going to require actually is that going to require custom silicon because it's like a different thermal footprint, right? >> Probably. Yeah, there are other there are different requirements in space from temperature from like kind of resistance of the chip and so on. So >> yeah, even like working on the Pixel phone, they have different corners. So like the phone has to be robust to different temperatures like different voltage settings and so like yeah, you just have to increase in many more corners I guess for these chips.

>> Yeah, that's a good example actually of like where the different properties of you know different heterogeneous compute you have different chips, different use cases. cutting chips in space right now like the whole data center is in space you know field is structured around how do you make existing uh GPUs perform actually in space but I think if everything for the pre-rade for your company works there's going to be specific compute for space >> customization for space in this case >> so you've already become incredible talent magnets in just the couple weeks since incorporating the company um we have Jiwoo Echen Dan, Iraham. Um, what what else are you looking for in terms of talent?

>> I mean, we're definitely building out on the technical side. So, every stage of tip design, we're looking for top talent, but also on the LLM side. So, this is like we're we're a frontier AI lab. So we want people all the way from like pre-training, mid-training, post-training, RL training, you know, experts in evaluations, data, and also on the non-technical side, we want, you know, operations specialists who can help unlock us like recruiting like um chief of staff, like that kind of thing.

>> What do you hope to have accomplished in a year? >> Oh, we're going to release our first product by a year, but yeah. [laughter] >> Yes. uh >> strong commercial partnerships. >> Yeah, strong partnership and we want to have our first product in that timeline that uh we can offer broadly not just to our partners but more broadly to other chip designers. >> Anything you're willing to share in terms of what that first product might look like?

>> Yes. Uh it's going to accelerate the process. It's going to tackle the long poles in chip design and it's going to be more end to end than the products that we have access to today. >> Yes. How do you see the role of a human engineer uh evolving in a world where the design process is automated? What do they end up doing instead with their time and talent? >> I mean I think there's like our company goes through various phases. So like the answer to that question maybe changes over time. But you know one way to think about it is like you know these engineers can come work with us and help us reimagine the process. So that's one way. [laughter] Um the other thing is like it's sort of like as humans start working with like claude code or cursor you know maybe they're not doing as much hands-on coding but they're becoming amplified and becoming much more productive through these AI tools. So something like that like for example even just alpha chip we could generate these like prito optimal curves of like different tradeoffs like every one of these points would have taken humans like weeks to generate but we could generate many many of them because it's it just takes so little time and compute for this model.

So like human can like explore like what is it we really want what trade-offs we care about. >> So we're going to vibe code chips. That's what I heard. [laughter] >> Maybe that's not quite but yeah maybe [gasps] kidding. So you've worked with some of the greatest of all time um Jeff Dean Noom Shazir many many more um very closely um Coach Lee and and others what are some of the lessons that you've learned from them or experiences that have really shaped you and and how have they shaped your perspective today? So I I think with uh all of these uh people that you met, they're extraordinary. So just setting the bar really high and um all of them are also extremely passionate about what they're doing and that's how Anna and I feel about this company like we are so excited to solve this problem no matter what. So that's something that we think is very important to actually uh be successful. And um the other thing about them is just they're just really nice people.

>> Yeah. >> And that setting that nice culture, collaborative c culture when where everyone is heard and uh everyone can do their best work. That's like a very important important thing that we want to uh follow their roots as well. >> That's right. I think like Jeff for example like incredible breath, incredible depth, incredible speed, but also like very high integrity. Like he treats every single person with like respect. It doesn't matter if they're like president of the universe versus like they're an intern or something.

Like it's equal like treatment. And I think that's like we want to embody that too. And I just remember like Jeff, he actually gave me like a performance review, a corrective feedback or something. He was like, "Ask for more from other people." And so I wanted to [laughter] Yeah, expect more from others, too. >> I like that. [laughter] I like what you shared. I I saw all these um photos of Jeff at Nurups running with like many people, [laughter] which I think perfectly encapsulates what you just said. He he welcomes others in and he you know motivates them and and spends time with anyone no matter what.

>> I remember clock also he was our manager almost for 10 years. He was telling us like deep learning makes all of us renaissance people like we can make contributions in many different types of fields and I think that's true here as well. And he also said like have fun in this company like actually enjoy. And I think we think that's important too. We think if everyone's having fun, then we're going to do even better work together.

>> I love that. >> All right. Thank you so much, Anna. We're so excited for your first year at Recursive, and we're really, really excited about it being the frontier lab for AI chip design. >> Thank you so much. [music] >> [music] [music]