**Hrant:** 0:00

At that time there were no good supervisors in our neighbourhood, so we started just to rent an apartment with a GPU and play with it. We wanted to make sure that there are interesting and relatively well-funded opportunities for those students who want to get deeper into some specific topic and actually get to the like world-class level of knowledge. Yeah, so we are trying to make it as a career, like an end-to-end career with, like, my own experience being a case study. But yeah, there are people who actually at some point decide that that's not what they want to do with their lives and then they go to industry. Some people go to PhDs in other countries and then go to industry. Our goal is to make sure that those who want to become professors have a good opportunity to do that. There are 200 people now in Armenia that are working in community rights and I think almost all of them have never really touched computers.

**Craig:** 1:03

Hi, I'm Craig Smith and this is Eye on AI. In this episode, I speak to Hrant Khachatrian, a notable figure in the field of machine learning and data science in Armenia, and he's the founder and director of Yerevann, that's Yerevann with two N’s at the end of the machine learning research lab. His research is mostly focused on deep learning for sequential data, including natural language and clinical time series. In addition to his role at Yerevann, he also works as a data scientist at IntelliNAR and he's contributed to the field of computational linguistics, with work published in the proceedings of the annual meeting of the Association for Computational Linguistics. Most importantly, Hrant gives us insight into the development of machine learning in Armenia, an amazing country that I recently vISIted. I hope you find the conversation as fascinating as I did.

**Craig:** 2:06

Hi, I wanted to jump in and give a shout out to our sponsor, Netsuite by Oracle. I'm a journalist and getting a single source of truth is nearly impossible. If you're a business owner, having a single source of truth is critical to running your operations. If this is you, you should know these three numbers 36,025, 1. 36,000, because that's the number of businesses that have upgraded to Netsuite by Oracle. Netsuite is the number one cloud financial system streamlining accounting, financial management, inventory, hr and more 25, because Netsuite turns 25 this year. That's 25 years of helping businesses do more with less, close their books in days, not weeks, and drive down costs. One because your business is one of a kind, so you get a customized solution for all of your KPIs in one efficient system with one source of truth: Manage risk, get reliable forecasts and improve margins everything you need all in one place. As I said, I'm not the most organized person in the world and there's real power to having all of the information in one place to make better decISIons. This is an unprecedented offer by Netsuite to make that possible Right now. Download Netsuite's popular KPI checklist, designed to give you consistently excellent performance, absolutely free at Netsuite.com. Slash Eye on ai. That's Eye on AI, E-y-e-o-n-a-i all run together. Go to Netsuitecom. Slash Eye on AI to get your own KPI checklist again. That's net suite.com. Slash Eye on AI. E Y E O N A I. They support us, so let's support them. So what? So we'll just talk. But why don't you start by introducing yourself, and then I'll just start asking from there. And it tells a little bit about your educational background and this lab that you're running or Institute, and that's her thing, yeah.

**Hrant:** 4:40

Great, yeah, so I'm Hrant. I have a PhD in computer science, in graph theory, from Yerevann State University. I studied here in Yerevann in this university both undergrad masters and PhD went into Computer science research From, I think, early years of undergrad. I had a great supervisor and got interested in my academic career. I worked in industry as well, so I was part of a company called Intel in air. That's a US based Company with Armenian roots that is doing computer vision for agriculture. Mostly operates in Midwest states, so at some point I was leading a computer vision team there. But then I realized that Staying both in academic life and in industry is not really possible. Yeah, so I had to choose and I chose the academic path. So With a group of friends we founded Yerevann, and, which is a nonprofit research lab. It's Yerevann, yeah, it's Yerevann, and so Yerevann is to the right of the town and double N is like a Neural network type of thing. We chose that name. I think it was in late 2015, when there were like lots of papers being published with something and then and then, so, yeah, yeah. So the goal was to create an environment where students can Work on academic projects in the field of machine learning. Mm-hmm at that time there were no good supervisors in our neighbourhood in Yerevann nearby, so you know, we started just to rent an apartment, hmm, by a GPU and play with it.

**Craig:** 6:49

Really? and so, first of all, you, who was funding this?

**Hrant:** 6:56

Yeah, so initially it was donations. Like even myself and other friends who Were in industry also, we were funding the initial operations. Later we started to find new sources, so initially it did not really require a lot of funding. Mm-hmm, all of us were students, and did not have many needs. But eventually all these problems started to become bigger and bigger and the big help that we got was from Aram Galastian, who was at that time he was a research director at the University of Southern California, ISI, so we started to work with them. The most important thing was that he started a Supervised project. He started to connect us to other researchers from his team and so on, and then we got into their projects that were funded by other sources, right, so we kind of started to subcontract the US projects and then government grants in Armenia started to pop up and we started to get those grants. We had several opportunities to work with some local other Foundations, nonprofits that were funding those type of capacity building stuff and more recently we started to also accept donations and there are people who actually donate to us and that enables to engage more students from young agent and connect them to the research projects that we're working on.

**Craig:** 8:32

Yeah, and, and that was what year that that you started?

**Hrant:** 8:37

We started in 2016. Like officially, we start 2016

**Craig:** 8:41

Right, right, right before the transformer craze.

**Hrant:** 8:45

Yeah, at that time there were no transformers or we had a project in NLP that was about extracting knowledge from biomedical literature. Hmm and None of the initial experiments were transformers. All of them were based on LSTM's rights. Yeah that's.

**Craig:** 9:04

That's fascinating and and so I mean I'm interested in how it's developing, just the capacity building, I mean how the community is developing. So when you started out, how many students did you have in the apartment? It's fascinating, you had one GPU. That's amazing.

**Hrant:** 9:23

Yeah, they started with like five people Mm-hmm, and then some sort of flow started, so one of them went to PhD in the US, then other students joined. I Think in two years we had the first case when someone from industry left his job and joined our team. So that was something that we did not know when and how it would happen. Right, because In Armenia the tech industry is growing very well and it's very well funded, so a lot of students that know something that can benefit Industry, they start to work at the very early stage, no, and to not develop like deeper skills in research and in like more deep topics, right. So that's one of the phenomena that we wanted to change with our approach. Right, we wanted to make sure that there are interesting and relatively well funded opportunities for those students who want to get deeper into some specific topic and actually get to the like world-class level of knowledge so that they could publish. In. That is not good conferences, good journals and so on right so this, so making this path as a feasible career path for a student that was an important goal for us and Somehow we can say that we achieved that. So now there are many students, including at masters level, even after masters, that are joining us Because they feel that it's possible to build a career in this way.

**Craig:** 11:13

Right, so so, build a career, meaning that that the at the lab, you call it a lab, or an institute lab at this moment, that that gives them the experience, that then they go into industry, or that gives them the experience and then they go into academia, or or is there enough funding that that's actually a career Doing research through the lab?

**Hrant:** 11:39

Yeah, so we are trying to make it as a career, like end-to-end career With, like my own, experience being a Case study right. Right, but yeah, there are people who actually at some point decide that that's not what they want to do with their lives and then they go to industry. Some people go to PhDs in other countries and then go to industry or or come back. So two of our alumni, the way we call them Are actually co-founders of tech startups in Armenia. Now, right, this also happens and I mean it's very natural. Right, like it's a funnel, and Most of the people do not End up being good professors. Yeah, but we, our goal is to make sure that those who want to become professors, they have a good opportunity to do that right,

**Craig:** 12:34

And you went to YerevaNN State University, which is the best school for, I guess, for AI. Is that right?

**Hrant:** 12:46

Yeah, but at that point it was not really like there were no programs or even subjects like on AI, so it had some traditions like more fundamental mathematics. So we started calculus and like mathematical analysis pretty well, which enabled All of this to happen right. So in that sense, yeah, it was good. But in terms of AI, it started like the first, I think, masters program that Our colleagues started. The theory on state was in maybe 2017 or 18, and now there is a relatively good masters program there, and then also a few other universities started, so nowadays there are several undergrad masters programs and also like private training centres that provide pretty good education in this field. So all of that happened essentially after we started

**Craig:** 13:44

right and? Are there labs, AI labs at the university?

**Hrant:** 13:49

Yeah, so the university started an AI lab last year and we are in the process of integrating in some ways, right, so the government strategy is to invest mostly in the university, not to like private initiatives, and there is this general understanding that we should join the efforts and we are in this process. I think, yeah, so we will be more affiliated with the university.

**Craig:** 14:13

Yeah, well, I think. And actually what got me interested? I was at the Digitech conference and you had a Sign on the wall that said Armenia needs more GPUs or something to that effect. Are universities? Do they have more compute resources at this point? And and talk about that, about how growing this community is is constrained by Hardware in many ways.

**Hrant:** 14:48

Yeah. So you know, initially, Like, if I start from the early years of starting all those activities, we were trying to position ourselves to as like less compute intensive research, less compute intensive work in AI, right so, and there are projects, there are directions that you can do with less resources and I personally as, like a leader of the lab, was trying to cut those projects that could grow into computing intensive stuff, right. And then this Article came out by Richard Sutton. We said that the bitter lesson, right. So, and at that point I Remember that I was very much against that, like I was like no, this is not true. We will prove with our example that it's not the case. You can do a lot of good stuff without large computers. And I think when GPT 3 Was announced and the paper came out and we started to read and I started to understand what's going on, I started to kind of doubt about how correct I was? And in, yeah, and I Started to realize slowly that it's getting increasingly hard To do relevant research without much computation. So the entire education that happens in Armenia in AI is also built around these non-computing, extensive approaches, right. So there are a few hundred people now in Armenia that are Working in machine learning the way we know it. We organise a conference every year called SDR 1 and we look at how many people come, and it doesn't make sense to come to this event if you don't understand machine learning because, like, the content is very Technical. So this year we had like more than 600 participants. So I think that's a relatively good estimate, because there are people who do not come when there are people who come and are not very relevant, but On average I think this is a good estimate on the size of the community right, and I think Almost all of them have never really touched a big computer.

**Craig:** 17:24

Hi. Good tech solves problems that you've thought about. Great tech solves problems that you haven't even thought of. What can the commerce platform trusted by millions of merchants do for you? It's time for Shopify, the commerce platform, revolutionizing millions of businesses worldwide. Whether you're a garage entrepreneur or IPO ready, Shopify is the only tool you need to start, run and grow your business without the struggle. Shopify puts you in control of every sales channel. So whether you're selling satin sheets from Shopify's in-person point of sales system or offering organic olive oil on Shopify's all-in-one e-commerce platform, you're covered. Shopify powers 10% of all e-commerce in the United States and Shopify's truly a global force, powering Allbirds, Rothes and Brooklyn and millions of other entrepreneurs of every size across over 170 countries. Plus, Shopify's award-winning help is there to support your success every step of the way. Sign up for a $1 per month trial period at Shopify.com. That's Shopify, that's EYEONAI. I'll run together. Give them a try. They support us, so let's support them.

**Hrant:** 18:57

Which is getting scary now, because you either become a user of OpenAI technology or you go to more computers and try to do something with more computers that could be competitive with those big ones. Again, you do not compete directly on chatbots. It's impossible to compete with ChatGPT if you don't have 20,000 GPUs. But there are many things that you can still do in research with 100 GPUs. If you don't have those 100 GPUs, you are slowly diverging into this situation where you become a user of the technologies that others develop with more compute. This is essentially the situation. We are in a stage where we have the capacity to utilise a lot more computers than is now available in Armenia. Getting back to the first part of your question, the university has access to scientific equipment grants. With those grants, we have bought a couple of nodes of NVIDIA DGX boxes, but we believe it's not enough. The science funding in Armenia works in a way that it's hard to invest several million into one direction. The general budget is limited. We got I don't know, half a million. We're looking for more options. We're working with the government, like other parts of the government, because it's not only about science, it's about enabling the next generation of AI companies that will have economic impact. You do need to do this capacity building now so that in two or three years you have the teams to compete in the business side of things. This is the strategy that we have right now. We're working with the government. We're trying to explain to everyone what's going on. What are the risks that the general AI community faces with the developments that open AI and Google are doing? The challenge is very general. I was at ICML this year. It was in Hawaii in the United States. I was asking many researchers how did ChatGPT change your research agenda? It's obvious that it's a big pain for everyone. One of them told me the research agenda is something else. I'm more concerned now about what I need to teach to students. They were teaching computational linguistics for decades. There are some people who are still in a kind of denial phase. They don't want to accept how many research directions are absolute. In some sense, this is a big problem for everyone. To stay relevant, even in academic circles, even in the US, if we have this decision to capture more of the AI value that is being created with all those technologies, we need to have these bold steps now to investing capacity, both in human capacity and in hardware capacity,

**Craig:** 23:06

Google and AWS and various, I'm sure, Azure. I don't know specific programs, but there are scholar academic programs that make computers available. Have you leveraged that? Also, how many GPUs do you think there are in Armenia today?

**Hrant:** 23:32

It's hard to say because most of it is in private companies. There are many companies in Armenia that have access to GPUs, but these GPUs are not physically in Armenia. That's also hard to count right. It's a methodological problem. For research purposes, the equivalents of A100s, which are the previous generation-based GPUs, I think there are less than 20 available for research for universities for these types of things. I think maybe around equally to that, maybe another 20 available to private companies. There are companies like Pixar who are investing a lot now for their local needs, but this is not directly accessible for students or researchers. They have their own problems. They have their own competition. Actually, Pixar does. I think one of those companies that understood the risks very early and started to invest very early. I think there is a good chance that they will stay relevant because

**Craig:** 24:53

NVIDIA has an R&D operation here now. Do they Do, does it do? Do they bring GPUs to the country and make them available?

**Hrant:** 25:04

Not really. You know, sometimes I have a feeling that even inside NVIDIA there is some kind of shortage. Hmm, so when I talked to NVIDIA folks at ICML by the way, we were discussing some of their new supercomputing architectures and new hardware and I was like do you have access to them? No, so this and the team in Armenia is mostly on the computer graphics side- yeah so they have access to GPUs. It's not the same. Essentially GPUs that are used for large-scale training of AI and video have a couple of teams internally, I think not very much connected to the Armenian office. That's two large language models, they had that megatron model and a few stuff like that. But yeah, it's not directly accessible to two researchers here and it's obvious, right like I mean, they have a lot of internal demand as well, so it's hard to yeah.

**Craig:** 26:16

Yeah, and so how many do you have at the lab now?

**Hrant:** 26:17

Yeah, so now we have access to eight A100s from your advanced at university. We have Like eight other GPUs of smaller scale that we do for smaller experimentation and we're waiting for another eight A100s that will come up fully in a couple of months. So we had again through government grants.

**Craig:** 26:42

Yeah, and what kind of research are you doing? I mean, we were talking last night and there's a lot of research right now in data pairing or trying to, you know, get more productivity out of fewer compute resources. So is that one of the directions you're involved in?

**Hrant:** 27:10

It is yeah, so for many years we were working again with our colleagues from USC. We worked on semi supervised, like object detection and things like that, right, and this year, a model from a meta segment and I think it came out and we realized that most of the research that we were doing is not relevant anymore. So this was again one of those small bitter lessons. Right so, they used a lot more computers and they achieved much better results and I think the past that we were following did not we would never come up with the same type of result, right so? And at that point we started to revisit our research agenda and we started to focus on two big directions. Mm-hmm. So, the first one is Language modelling for essentially for drug discovery for molecules, for small molecules that can become drugs can cure diseases, right. So there is some public data available about them. Biologists do not really trust that data. There are many funny stories about how unreliable that data is, but we are confident that there is something to learn from that data. So again, we kind of apply those language modelling ideas To that so we encode these molecules as text and code this data as text so that we could kind of leverage Transformers and then other good things that happened that were developed over these years, and that project will go towards interactions with proteins, like more biological effects of the molecules. Now we are a little bit more focused on kind of simpler things like chemical properties, not directly Attacking the biological aspects of it. Then on the roadmap we have like the synthesis problem, because turns out that people can generate molecules in the computer and do some simulations and show that it's good, but then you go to chemists, you ask them to create this molecule and they say you know, I cannot synthesise this or Doesn't make sense, right. So that's another interesting challenge that there are some computational approaches right now that people use, but the power of language models has yet to be demonstrated in this field as well, right? So that's one general direction. I think we have like five, six, seven people working Generally on this direction

**Craig:** 30:13

And just before you go to the next research direction, what models are you using? Are you using open-source models?

**Hrant:** 30:19

Yeah. So, it's an interesting question whether you need to start from an open-source model or you can start from scratch, depending on the size of the data. So, we are also examining data aspect now. Right now, we are starting from metas models like Galactica, which was trained on scientific texts, but we don't have a clear demonstration that it really transfers this knowledge to the new data right. We're in the process of evaluating right now.

**Craig:** 30:55

yeah, and then the other research

**Hrant:** 31:58

yeah, so the other direction is More related to vision, computer vision. It's more like a compute constrained foundation model for mostly aerial navigation. Like you are a drone and you want to navigate some environment, maybe you are Very high, you use satellite data as well, maybe you're flying low and you need to I don't know to avoid obstacles. So, you need to be able to understand your environment with not a very big model, but not also very small, right? So, we're trying to see what the balance is there. You also want to capture many modalities with many cameras. You don't want to rely only on RGB, you also want to understand infrared and maybe some Radiometric data like radars, whatever. So, this is the general direction and also joining it to control related stuff, like not the control at a lower level, but like decision-making on what action should I make right now? Right? Like not at the level of how fast should I spin my one of the motors, but like on a higher level, like what would a human drone operator do? Right, I don't know, use a joystick to navigate, right, like? Can these models understand the instructions, like the mission, and understand the environment and make a decision whether I need to go left or right? Right? So, this type of model. So, we were designing this like a half a year ago and I think in July there was a great paper from DeepMind. It's called RT2. I think it's a way to leverage language models to perform actions of a robotic arm. So it's a robotic arm on the table. There is a data set that some people actually annotated it. So you give it in instruction, like I don't know, move this bottle to this computer and it actually uses Robotic arm to perform the action and you record, like what it did, right, so Rotated this part of hand, this many decrease and yeah, and then you give this to language model. So, the language model has a camera. I see it's. We call it a language model, but it's obviously beyond text, right? So yeah, it has a text instruction, it has an image and the output is another text that is interpreted as an action. On how many degrees you should rotate this right? And then they show that if you start from a huge language model that they have internally, it can bring world model, world knowledge in some sense. So, you ask it to. I don't know, push this bottle to Germany and you put flags of many countries on the table and it pushes towards the flag of Germany because it knows what is Germany, what is flag, how it looks like it has read the entire Internet right before. So, it kind of brings this interesting effect that you can Transfer this general proposed knowledge into specific robotic tasks. Hmm. Right. So now this is exactly what we wanted to do and, yeah, we are In some sense getting behind because, like, obviously, this guy's with a lot more compute and a lot more resources are faster, but on the other hand, the field is very wide, so there are many things to do there and their models are not open and are huge, so there is no chance you can run on All the GPUs that we have in Armenia, these models, and they are like I don't know 500 billion parameter models. So, I think, yeah, that's another big Direction that we're working on.

**Craig:** 35:17

Yeah, on that. I mean that's, that's interesting, that work is being done.

Where did you say deep mine in your mind? Yeah, and for you. Oh what? What language, language model are you, are you using?

**Hrant:** 35:36

So the early experiments that we did on that were based on it's called open flamingo. Flamingo was another deep mind model, right closed source, and I think a team from University of Washington started to replicate that using open data and yeah, we started from there. I think we will switch to more powerful things. Looks like there are a few other things that became available over the past couple of months, but yeah,

**Craig:** 36:08

I mean you, and when you say others, you mean closed models that you access through APIs or

**Hrant:** 36:14

No, no, it's those things are. I think at this point they are not. It's not possible to access them through APIs, because you need to fine tune them. Yeah. So what we're doing, we're taking open plumbing or something else and we are fine turning them on the robotic or drone related data. Some of that is available. Actually, recently, deepmind, with many, many partners, maybe like 30 partners released a large data set for table based robotic stuff. So very different robots, very different tasks, and they are trying to build a unified model for all of them, like one model for all robots, and a lot of data is available there and we are also looking for actually collecting our own data, maybe our own sensors, our own missions, like, for example, agriculture is pretty interesting, like go this way and find the tree that does not look good and maybe spray something on the tree. Oh, I don't know, count the number of apples in the field, right? So there are these types of missions that you can plan and someone like an experienced drone operator can fly and generate data and then you can try to teach those models to perform the actions.

**Craig:** 37:42

Yeah, on the robotic arm thing, the control mechanism is AI as well.

**Hrant:** 37:53

So I think it's at the level of go left, go right, so, or at the joystick level. You would push this button, so the AI pushes the button, but whatever happens underneath it, the underlying control system takes care. So, actually, I think it will be one of the interesting research questions, like down to which level you would want to cover with AI and then pass to the more traditional control. I think the level that we are looking at right now is this push the button level, but who knows? I think there is a possibility that it will go beyond that. When you go beyond that, you have speed concerns. You have to be very, very fast, which brings all these questions about the model size and everything. And then there is the safety side. Yeah, so a lot of these control algorithms that are employed by these robots or drones have some safety guarantees that can be somehow mathematically proven. When you go to AI, you will have challenges there. I know it's an open research direction. We are not looking into that right now,

**Craig:** 39:20

Do you then vectorize the movements so that you could use a large language model, a transformer based model, to give the instructions to the

**Hrant:** 39:39

yeah, you can vectorize or tokenize everything. Yeah, right, yeah. So the way people do right now with the ChatGPT systems, every action can be kind of expressed as a function call, right? You can say like this is a function name, and then you have some arguments which are probably numbers or something, right, and this entire thing can be expressed as a string, as a text Character by character output, the function name and the numbers like digit by digit, right, this might sound stupid, like you could do more specific things, but if you're operating in the language model, mindset right this is a very reasonable thing to do, so DeepMind was doing that. So to get into more specifics, a robotic arm is a six-axis object, so you control it by deciding six numbers. So their language model explicitly could output those six numbers in text. So that's it. Yeah, I have seen papers that try to get more clever, like the way you tokenize the numbers, is non-trivial, like in GPT-like models, do you? Just, I don't know, every digit is a token, or maybe a couple of digits, pairs of digits, are tokens. Some papers try to do very clever things, like I don't know if the digit is right next to the dot it has a different embedding than the digit that is further, like if it's in decimals it's one thing, and so on. I personally believe that you really don't need to do these things if you have a big enough model, because those are like very simple mathematical tricks that transformers will be able to learn easily, right? But I cannot be sure about smaller models, like if you really want to squeeze your model into a very energy-efficient thing, maybe those clever tricks are still useful, right? It's obvious that in the big models these are not like I'm sure ChatGPT does not use any specific tokenization for any specific symbol or context Right, so it all works with the same basic principles, yeah,

**Craig:** 42:27

So how do you feel about it? I mean, certainly, the communities have developed tremendously since 2016, did you say yeah, yeah, and the research community is very open and information passes very quickly? Do you feel that you're part of the global research community or do you feel isolated in Armenia? And as you build the human capital here, do you feel like Armenia will has an opportunity to push the? I mean, right now you're I don't know how to put this but you're not at the leading edge. You're sort of following research trends, but is there an ambition? I would imagine there is, that then you could rise to the leading edge in some areas.

**Hrant:** 43:40

Absolutely yeah. So I think the international research community in AI is open to the extent that, no matter where you are, you have the opportunity to be connected and not feel isolated. And when I say you have the opportunity, it doesn't mean that everyone is using those opportunities right, so you could easily get isolated as well, but the opportunities are there and we spend a lot of effort to stay connected and I think we are relatively successful in that. So the way we do this is to make sure that almost everyone in the team attends one international conference every year, right, no matter how far they are. The other way is to invite well-known people in the field to Armenia. So we have this data FSDR one, which is a joint effort with our colleagues from industry. So it's not like an academic event, it's more like more focused towards industry, but we bring scientists as well. So every year, like 10, 15 people come and we do like networking and conversation. So we're trying to make sure that this flow exists in both directions so that whoever wants to be connected and not get isolated they have the opportunity. Another way we do this is to raise the bar of the work that we're doing Right. So in our team. We, like everyone, know that if you don't publish in a top machine learning conference, it's not like you can't say you did your work right. So everything you do has to be published in a good place. You should be put on GitHub so that others know, right. Otherwise, like, that's the criteria for success. So this is the second thing that, like, if you come to our office, you know we have a list of journals and conferences ranked by Google Scholar written on the wall so that everyone, every day can see what are the good ones and what we want to target right. The other thing that is very helpful in the case of Armenia is the diaspora. So we have many Armenians that are publishing in those top places and attending these conferences and working on this research, both in academic environments but also in companies like Meta, google. So we have a very good relationship with them. They help us a lot, they engage us in some of their research projects. Even so, we have, like this year we had, I think, three papers with different ratios of contributions with Meta. So our colleague there, Armen Aghajanian, started this and then it grew into larger collaborations with more people. I think he had four papers, yeah, and the fifth one will be sent, maybe in a month. So all of this requires a very high bar of quality, because those folks are not interested in low quality research or mediocre research, right? So I think it's very important that even the undergrad students who come to our environment, they start to feel the pressure of staying on top of things in general. So we have papers in the top conferences, like a couple of them every year, but that's not the level that we are aiming for. So we're aiming for more impactful and more involvement in those top venues. So you know, initially our research agenda was more, more governed by those supervisors that we were able to find in different places. So we were doing that and we could work on 10 projects with 10 people, right, because there is one supervisor who can lead one student and yeah, that's the early stage of capacity building. Right Now we are at the stage where we are designing our research agenda and the two big directions that I mentioned are already kind of our own products, right? And again, obviously, we consult with all our colleagues from here and there Armenians, non Armenians, everyone we can get access to. Yeah, the ambition right now is to become a centre where we produce high quality research regularly in these two directions for many years. Right, so that we secure our place on the map of this direction. So we were careful not to include, like, a lot of NLP in that, because it's very, very hard to compete there, right? So these decISIons, like why we chose chemistry and this aerial navigation stuff, is based on many, many factors like our background, our connections, the competition In research it's not really competition, but you know what I'm talking about so our chances to be competitive, essentially our resources, current resources and perspective right. What we can expect in the next couple of years. Yeah, and also the way we see where Armenian industry can go right. For me, it's obvious that the big part of AI, the big part of the value that AI will create in the world, will be in biotech, and I believe it's strategic to develop capacity in this direction. So that's another important factor why we got stuck in this direction. And also, like aerial based everything is also important for many reasons, like including agriculture, like disaster handling right.

**Craig:** 50:54

Yeah, I don't have much time left. Actually I'm sort of, but let me ask, we were talking last night, I mean. So you're constrained by hardware and the current generative of AI or transformer based research is very compute intense. But there are other research areas that I think are promising and personally I think have a better chance of reaching artificial general intelligence or that you know, which is kind of the ultimate goal for a lot of researchers, and I'm thinking we talked last night about Yann Lecun JEPA models and Rich Sutton's work you mentioned Sutton in reinforcement learning. Why not, given the hardware constraints, why not pick those areas where, frankly, it's not as crowded, there aren't as many people doing research, doesn't require as much compute, and I would think there's a better chance of having a breakthrough or doing something significant.

**Hrant:** 52:28

Yeah, good question. So, like, in terms of reinforcement learning, we don't do reinforcement learning, it's not in our background, like we cannot do a shift like that and personally I'm not convinced, to be fair, that it's a better approach to solve many problems. Right, I'm even sceptical about this recent RLHF stuff that is being built on top of language models. Like I believe that the pure pre-training of language models is a lot more elegant and beautiful and powerful, and the RLHF part sometimes has this feeling. I have this feeling that it is a little bit of a hack to make language models I don't know more user friendly, so that people do not complain or I don't know. So the general learning that happens inside the best models we have right now is actually happening in the pre training phase. So I believe that the big challenge in AI research now is to actually extract this information from the big models. I think we're not doing that well enough. So these models actually know more than we are able to expose with the strategies we have. I'm sure reinforcement learning based solutions can be part of the solution. We are more interested in other aspects. So one paper that we have submitted recently was about looking at vision transformers which are like not that big, like a couple of billion parameters maximum, and the smaller versions of them, and we're trying to see how much knowledge they already have about parts of the image and how much you can extract from it without any further training. And it's an interesting set of opportunities. What can you extract and what are the tools to extract? How do you deal with it? It's almost like a psychological process, like how do you convince the model to expose the knowledge that it has? So these types of things are very interesting. So I think that, to directly answer your question, I think that the way which models capture knowledge is unbeatable now and I'm not sure that other approaches will be better. I appreciate the work that people do outside of it. It's not that obviously you should not put all the X right in the same place, so people should work on other things. Our approach is to stick with these approaches to see, to try to understand how much those transformers could extract and how we can expose it and how we can utilise it, and then do this for chemistry, like there are no chemistry GPTs now right, so let's see if we put the whole publicly available chemistry related data to these transformers. What will be the model and then what we can extract from it, like which biological important problems in the drug discovery pipeline can be accelerated if you have this model right. It's not really obvious how to do that and what can be extracted there, and I think at some point we will just release the model to the public and see what other people can do that we didn't think of. So that's the strategy we are taking. So part of it is computer intensive but a lot of it is not computer intensive because you are playing with the existing model. Yeah. I remember in one of the interviews, Elias was clever also, so there was a question like how much time do you spend on generating new ideas and how much time you spend on implementing those ideas, and he was like. You know, like 90% of the time we spend is to understand what these models are capable of. Yeah, yeah. That's right. When I heard that, I realized that a lot of the work that we are doing is the same. Many of our papers, like last year we published at CVPR, were about failure modes of domain generalisation, like if your model was trained on some distribution of data and then the distribution is changed. How your models fail there are. We discovered that there are many different ways to fail.

**Craig:** 57:57

All right.

**Hrant:** 57:58

So you look at the numbers. All numbers are bad, but then you realize that in the details there are many. They are very different. Yeah. Right. So this approach is pretty interesting and exciting for us and I think a lot of our research will be on those lines.

**Craig:** 58:19

Hi, I wanted to jump in and give a shout out to our sponsor, Netsuite by Oracle. I'm a journalist and getting a single source of truth is nearly impossible. If you're a business owner, having a single source of truth is critical to running your operations. If this is you, you should know these three numbers 36,000, 25, 1. 36,000 because that's the number of businesses that have upgraded to Netsuite by Oracle. Netsuite is the number one cloud financial system streamlining accounting, financial management, inventory, hr and more. 25 because Netsuite turns 25 this year. That's 25 years of helping businesses do more with less, close their books in days, not weeks, and drive down costs. One because your business is one of a kind, so you get a customized solution for all of your KPIs in one efficient system with one source of truth: Manage risk, get reliable, forecast and improve margins everything you need all in one place. As I said, I'm not the most organized person in the world, and there's real power to having all of the information in one place to make better decISIons. This is an unprecedented offer by Netsuite to make that possible Right now, download Netsuite's popular KPI checklist, designed to give you consistently excellent performance, absolutely free at Netsuite.com. Slash I on AI. That's I on AI, e-y-e-o-n-a-i all run together. Go to Netsuite.com, slash I on AI, to get your own KPI checklist. Again, that's Netsuite.com, slash I on AI, e-y-e-o-n-a-i. They support us, so let's support them. That's it for this episode. I want to thank Herant for his time. If you want to read a transcript of today's conversation, you can find one, as always, on our website. I'm on AI, that's E-Y-E-O-N-A-I. In the meantime, remember the singularity may not be near, but AI is already changing your world, so pay attention.