Nathan: Anton Troynikov welcome to the Cognitive Revolution

Anton: Glad to be here.

Nathan: So tell me first about vector databases. I think, you know, our audience is obviously interested in ai. They most probably have at least a superficial sense of what a similarity score is, what a DOT product is. But take me to how that works as a database and then, you know, we'll get into some of the optimizations there as well.

Anton: Yeah, sounds great. So, you know, let's kind of start at the beginning. What a Vector database allows you to do is if you can represent your data as a vector, it makes it possible to perform geometric operations on your data. So traditional databases have, let's say, algebraic discrete data structures, right?

Anton: Like a, a sequel table. You select from where, and it has concrete properties and it returns exactly that result. With a vector database data is represented as vectors in high dimensional space. So you can perform operations that have geometric meaning instead, like, you know, distance, you already mentioned similarity scores.

Anton: You can talk about densities meaningfully. You can, you know, fit surfaces to them, you can cluster them in a way, you can't in a, in an algebraic form data structure, discrete data structure. So what I think the reason that they're exciting right now is because we now have powerful models that can transform all kinds of different types of data into vectors, and it can transform them into vectors in a meaningful way.

Anton: So traditionally that's been text and a lot of these embedding functions have been kind of hand rolled. But now with the success of, you know, transformer models and attention based models for, for modeling text, The vectors that you get out from those kinds of models are much more meaningful. They, they represent structure like a, like a human would impose structure on the resulting vectors.

Anton: So, you know, sentences that, that talk about similar things, they end up close together in vector space. And that's kind of what that's done is turn on a lot of experimentation and development with using embeddings as kind of the AI native way to represent data. And so, I'll give you an example, A typical workflow that you see is you want a model to answer questions about specific information.

Anton: So chatGPT for example, it's trained on an enormous amount of data. You can ask a questions about anything. But it might hallucinate, it might not have the information that you want. And certainly it might not have your personal private information that it didn't, you know, it wasn't trained on because it's not available anywhere else.

Anton: But you still want it to be able to answer questions about it. It's good at general question answering as a task. So what you can do is you take the corpus that you want it to answer questions about. And then you embed it.

Anton: You can embed it using, there's a bunch of open source models. You can also use Open AI's API, you can use APIand transform that data into vectors.

Anton: And so you have the data available as vectors with the actual text associated with it, with that representation. And then what you do is you query it and you query it by, again, taking a query, embedding the query text, performing a nearest neighbor searcher, an approximate nearest neighbor search in the in the vector database, returning relevant documents.

Anton: And then what you do is you take your question and you take the relevant documents that it found, and then you ask the model specifically to answer the question with the documents that you've put in its context. And then it's much more accurate and it's answering questions about the thing you actually wanted to answer questions about with data that wasn't available to it at training time.

Anton: And this is a really powerful technique. It pretty much suppresses hallucination. It allows you as an organization to use large language models without fine tuning them, which is expensive and slower. And it confers a lot of other advantages as well, which, which maybe we'll get into in time. But that's, that's kinda the basic idea.

Anton: Vector databases have been around for a while, pretty much any service that you use that has a recommendation or like a similarity thing. So the one that I always think of as Pinterest, it'll show you similar images. And the way that that's being done is because they have like these images encoded into vectors and they've got a large vector database so that like when it wants to find nearest neighbors, it'll, it'll look for that.

Anton: And you get visually similar images because they're embedding function produces visually similar images for vectors that are close together. But the use cases for these things are exploding now because the embeddings for all

kinds of data became available and it's super cheap and you can just hit open AI's model api, you don't need your own model anymore. And people are building all kinds of really cool applications on top of this stuff.

Nathan: Okay. There's a lot to dig in there and, and follow up on for one thing, you know, just to contextualize a little bit, I would imagine a lot of our listeners have experienced querying SQL databases. So I would love to get a little bit of a sense for kind how the, you know, DBA life is different as we move into a vector world.

Nathan: I think you, you spoke, and I'll just give a couple, you know, additional notes too on like, there is so much energy around these kinda chained retrieval and language model systems right now that's really at the heart of kind of these dreams that everybody is, is dreaming. Everybody sees ChatGPT and everybody says the same thing that you said, which is, well man, if this had my data, you know, if this could go read all my emails, if this could remember every text message I've ever sent.

Nathan: And then obviously, you know, the business use case. Is even bigger and probably a big, you know, a deeper need because now you've got organizations and all of our Google docs, you know, and our whole know, I mean, how many companies have rolled out these knowledge bases or the wikis and, you know, but do those things, can anybody even locate the information in there?

Nathan: Generally not. Right. So this vector database kind of sits at a really critical point in that value chain, and you've started a company that is a Vector database company. So tell us, before we go into even more details, tell us a little bit about your company specifically.

Anton: Absolutely. So, I wanna say one specific thing here. Chroma uses a Vector database as a technology, but to think Croma really is, is a platform for basically embedding knowledge that your machine learning enabled applications can use. You can think of us as the thing that acts as the storage layer for applications that use large language models in the loop.

Anton: So we're much more than a vector store. And one of the things that we have to really learn while we were building this product was the affordances of the existing vector stores on the market don't really suit this use case very well at all. And we actually built this product in-house and we were using it in our own product experiments for a while.

Anton: Chroma basically started on the principle that you can understand the model's behavior based on how it sees its data basically. And, and embedding is a representation, is a model's representation of its input data.And so our early, all of our product experimentation was around like running algorithms and embeddings, representing them, showing them in different ways, manipulating them in different ways.

Anton: And we needed something really easy to use for developers. We needed something that would allow us to rapidly experiment, maintain development velocity while still being very performant. And so when it became clear and, you know, OpenAI did a lot of good work here in marketing the power of embeddings to people.

Anton: And these sort of early use cases showed up with document question answering. It became clear that wait a minute, you know, there's thousands or hundreds of thousands of developers in the world that are gonna need something exactly like this. And we essentially took what we built, packaged it up to make it a lot easier to sort of just get up and running and then gave it to people.

Anton: And people really seemed to love it. Things have been really exploding over those. I mean, it's kind of crazy to me cuz it feels like it's been out for a while, but we launched it last Tuesday was when I announced it on Twitter.

Nathan: It's kinda been a lot since then. Yeah. Time is doing very weird things for me as well. I've, earlier this week I realized it was on Wednesday that I realized, wait, we just did that on Tuesday . And it was like, oh my God. So many things happened in such a small period of time in the AI world. It is dizzying to say the least.

Anton: I think, I feel like there's been a few step changes. Right Over the last couple years, I think I, so first of all, I think GPT 3 was the watershed back in 2020 and then Stable Diffusion DallE. But stable diffusion was the next really big one because it allowed people to see that, wait a minute, you don't actually need to own a huge model. I can totally just run this on my laptop and get results and build things around it.

Anton: That was a huge watershed moment because it was open source and, and like small enough to run on an m1 back. You can do it yourself. Someone was crazy enough to get it to run on an iPhone, which I thought was amazing. And then the next thing that just recently happened was chatGPT and technologists will all tell you the same thing.

Anton: People have been like following closely for a while, will tell you the same thing. Be like, oh yeah, the model was already available. It just needed fine tuning. Nobody did it blah, blah, blah, blah.

Anton: It doesn't matter. What matters is it's packaged and given the right affordances so that people can easily understand what it's, they can do with it. And then they'll start doing that, and then they'll start experimenting. So I think those are the two big watershed moments that happened in the last, like while that really makes it look like a amazing playground for experimentation right now.

Nathan: So you mentioned a couple things that I just wanted to elaborate on a bit. One, it sounds like you guys are a neutral database, vector database technology. So I can use any different provider. You mentioned open ai, their pricing, just to give context on exactly how cheap it's, cause you said it's cheap cost to embed all the transcripts of every podcast that Joe Rogan has done.

Nathan: Hundred pages of text. $40. Thats hundred thousand pages. And if you wanted to go --this is from Boris Power who 's at Open AI -- if you wanted to go to the full web scale data set of 10 to the 11th pages , then you're looking at only $50 million to embed. So you know, that's not nothing. But to embed the entire, you know, webscale databases actually is amazingly cheap.

Nathan: So once you have a a document, you know, or whatever it is embedded, then you can also store those embeddings. You don't have to, unlike the other APIs, you know, that are generative and doing something new each time. The embedding is kinda of a one shot. You know, there's no temperature setting. You're gonna get the same thing out every time.

Nathan: So then you can store it and that's where your platform comes in. You mentioned that there are some shortcomings of the things that were out there. So what shortcomings, did you see, and how did you address those in the product that you built?

Anton: Yeah, those are great questions.

Anton: I think the first shortcoming was just how hard it was to get started with anything out there. We evaluated literally everything else in the market and it just didn't fit our needs of like rapid development, rapid experimentation. It was either too hard to like deploy and keep running because it was really designed for heavyweight workflows from day one, or it like, didn't really provide the kinda price performance point that we needed.

Anton: Or it was just like, frankly, too complex. Chroma for a long time was just me and my co-founder, Jeff Huber. And we couldn't spend like an entire full-time engineer's worth of work just running the vector store part. We needed something else. And the other thing is, is we found the abstractions were pretty wrong.

Anton: There was also like a bunch of missing features, which, which we just couldn't figure out why they were missing. We just implemented them. There was a bunch of other things too, which, which kind of made it. Obvious that the affords here were wrong. Like none of the other products on the market are really like AI native.

Anton: That's the way we think about it. And that's because like you need to, like if you have this app that's LLM enabled and it has like a store, you need a much more like transactional approach. You need good inserts, updates, leads. You need to, you know, you need to have it talk to the embeddings API properly.

Anton: You need to have it return the results in a way that the model is gonna pick up easily and then stuff them into the context window. All of these things. And we also saw power in parallel the growth of what we think it was application development frameworks for ai, LLMs our friends Lang chain, for example, where we were a day one integration partner when Chroma launched with them.

Anton: Like we saw that, okay, the thing that we build has to actually fit what people are doing. It doesn't, it doesn't, it shouldn't fit this like, abstract idea of vector database. It needs to be an app development platform. It needs to be the storage layer for app development. And, and, and then it needs to have the right affordances to think about this so that developers can actually think about it in the right way.

Anton: Those are kind of the main differences. That said like it's pretty performant making it go even faster pretty soon.

Nathan: Pre AI, what were the main use cases for vector databases that people had built all these other things for?

Anton: Yeah, look the two big ones, as I mentioned earlier, are recommender systems and semantic search. Prior to embeddings, embeddings are a concept that basically just means take data for one high dimensional space and encode it as a vector in a lower dimensional space. That's all embeddings really means.

It's actually not even a deep learning concept. It predates it by a long time. It's an idea in mathematics, which, you know, the name is like loosely used now, but that's pretty much what it means.

Anton: And so there were a lot of, like I mentioned, especially for natural language processing, like hand tuned methods to convert words into vectors or convert like corpuses of documents of text into vectors. And then, you know, you would try to hand engineer these embedding functions and then store them in a vector database and then query them and look them up in a similar way to the way I've described.

Anton: And what people focused on in that era was building these really large scale data, like vector databases is intended to work at very large scale. So we're talking about like finding, you know, being able to query billions of documents outta the box, being able to like search over billions and billions of images out of the box for sort of these like web scale huge applications.And at that time, like AI just wasn't in the picture. Deep learning really wasn't in the picture.

Nathan: so do you also have like a sort of sql-like aspect to the database where, you know, you have a could I, could I do like a sort of equivalent of like a wear clause where I'm like, but I'm only interested in these.

Nathan: So you have kinda traditional B-Tree indexes along with the vector space.

Anton: Ok. That's right, that's right. It allows us to, so we have an underlying algorithm that we use for the vector index. And the index is the thing that actually stores the vectors and the relationships between them. And then there's a meta data store and the meta data store stores the data in like original, more sql form.

Anton: And we have them talking to each other in such a way that you can take something that filters the query that you perform in this very traditional way and then turns that into a query onto the vector store. So I can say, you gonna find me all the matches to this query, but only from documents that contain this phrase. And like, you can just do that with Chroma.

Nathan: So one thing you said about the technology that caught my ear was the, was the phrase approximate nearest neighbor search. So, tell me about that, that sounds and this also starts to preview a little bit the, the project that you've released and made some noise on the internet with, which is stable attribution.

Nathan: And we can get into that in a little bit more detail. But you know, the big thing there is you have a ton of images, right? I don't know how many of the of the images you ultimately, so you, you embedded all 5 billion.

Anton: We had that with Clip and Orbit. Cliff. And we can get into that one. We discussed that project. I do wanna come back to your note about approximate versus exact nearest neighbor search.

Anton: So, Exact nearest neighbor search is, has like n squared complexity.

Anton: Each time you add a new vector and you wanna query it to find out what the nearest neighbors are, you need to perform an n squared calculation. Like it grows in the number of vectors, quadratically, which is actually for certain, for certain sizes of data, is totally fine. Roughly if you have like, I don't know, a couple of hundred data points is fine, you should just be doing exact nearest neighbor.

Anton: It's just one matrix multiplication and that's very fast. NumPy calls out to like very fast numerical libraries. You can just do a matrix, multiplication, keep it all in memory. It's nice and fast, you don't have to think about it. The issue of course, is Quadratic growth is really bad. So by the time you have, you know, north of 10,000 data points, that n squared is really starting to hurt you.

Anton: And this is where approximate nearest neighbor comes in. And there are several of these algorithms, but basically all approximate nearest neighbor algorithms are flat in the number of or approximately flat in the number of vectors in the index. So you no longer have this N-squared term. Your query times, regardless of how many vectors stored in the index are going be identical.

Anton: But there's a trade off. The trade-off is you won't necessarily get all the nearest neighbors every time. You might lose some. And that's a, that's called a precision recall trade off. And it's very similar to other aspects of machine learning in that way. But in practice, in many applications like these, that's a trade off that you can make.

Anton: And the trade off isn't very severe. It's not like you're gonna lose half, it's more like you might lose one in every 100 queries.

Nathan: Yeah. Interesting. Okay, so that makes a ton of sense. If you're working at a 5 billion vector scale, how do you think about kind of, you know, when you said 10,000, I was thinking, I'm pretty sure I have more emails in my Gmail than that.

Nathan: So what would, what, how should I think about kind of these sort of mid-scale personal data sets where I probably do want the full thing, right? Like I don't wanna miss the one thing that I'm, that I most need. So yeah, tell me in that mid-scale what happens.

Anton: Yeah, so you can, especially for smaller data sets, you can tune for more recall.

Anton: You can make it very unlikely that you'll miss anything but still get that good query performance. You can also trade off processing upfront and sort of insert, you know, time to insert something with recall and query performance. Depending on what your workload looks like. Basically, if you have a lot of inserts, if you want it to be fast, it's probably a good idea to, to, you know, tune it to be a little faster.

Anton: So that insertion speed, little bit rest. Recall, if you're like, insert, if you have like a set of data and you really need to insert it once, do all the pre-processing on that side, you'll still get great recall.

Nathan: So the tuning, it's interesting the. There's multiple meangs of tuning. You're speaking about a pretty classic database tuning paradigm, which is like, when exactly do we do stuff and do we pre-calculate when?

Nathan: So I had listened to your conversation with Packy on the Not Boring podcast not too long ago, and actually as I was listening to that, I thought you were tuning some sort of learnable you know, like fine tuning some sort of aspect of either the embedding or you know, the way like, which, so you're doing that as well, like whi which dimensions, you know, would be kinda primary. So tell, okay. So tell us about that side as well.

Anton: Yeah, so this is the other, this is the, again, like I mentioned, one of the most important parts is you can perform geo geometric operations and because of the way that embeddings work, those geometric operations have meaning. So for example Kind of the simplest example of this that I can think of is you can cluster vectors geometrically, but those clusters will have semantic meaning.

Anton: So if you have a large corpus of text and you look at what the clusters might be in it, and you look at the topics that are in those clusters, that's information that like belongs together because the model is trained to keep things that seem together. Together, right? Clustering is like this basic first approach thing you can do.

Anton: Another thing you can do, and this is actually a great for question answering application. We've been examining this in some detail lately. You can find a direction in the vector space that represents some human understandable property, and it's actually surprising how often this comes up. But for example, you could be talking about a topic, but you could be talking about it in the context of a question like who was born in year, blah, blah, blah.

Anton: And then over here you could be talking about it like, oh, this king, et cetera was born in blah, blah, blah. And so if you find this like question answer vector, you can do interesting like remapping and bring things closer together. And because because it's just a vector space, it's this well understood, fairly straightforward mathematical structure.

Anton: The way to transform one thing that you want into a different thing that you want or like make it work better, fine tune it to get better results for you is actually very straightforward compared to traditional machine learning. You don't need to do back prop, you just need to estimate some transforms, which is something that we know how to do.

Nathan: That sounds pretty similar in some ways to like classify your guidance almost, right? Like you're sort of identifying a direction that is the, the direction from statement to question essentially. And then you can kind of move around, you know, you can, you can place that vector anywhere in theory and kinda move from different, you know, statements to their corresponding questions.

Nathan: Obviously it's gonna be a little messier than that.

Anton: But there's also all kinds of other things you can do directly by working with this. For example, if you have a system to get user feedback about relevance of the return results, you can reweight your, your space by applying a learned transform on top of it to get ever more relevant results from human feedback in a very straightforward way.

Anton: There's no model retraining that you need to do. Your model is doing what it's best at, which is like understanding and recomposing knowledge. And actually that, that takes me on an interesting tangent here, which I really do wanna indulge briefly. Right now we're seeing like, oh, chatGPT, and then

Anton: GPT 3.5 and GPT four soon, which are these large models that know things and like they're trained on very large corpus of data.

Anton: They're very expensive to train. They're expensive to fine tune. You don't really know what knowledge is in them, but they're very good at general purpose tasks. They can do all task efficient for you.

Anton: And I think those models are gonna be around I think they'll continue to be there because they're such great, like, they're almost like utilities really. But on the other side, I think what we're gonna see is these smaller, leaner models, which are actually trained to find and compose knowledge in response to queries rather than store knowledge in their own weights.

Anton: And we've seen early signs of that. There's a paper called Retro or there there's a system called Retro. It's from Meta or is it from Meta? I think it's from Deep Mind. Sorry, I misspoke. Which where they trained the model to do exactly the sort of stuff we're talking about. They trained it instead of sort of storing in knowledge in its weights, they trained it in such a way that it has to search for knowledge in this data bank.

Anton: It can't get knowledge from anywhere else except in this data bank. And that, and, you know, that's, I think that's one direction in the future. And we're starting to see the very first stages of that happening right now. And, and the sort of stuff we're building with chroma really enables that. So those transforms that you described, those are just like one layer, like you're, you're basically just learning one additional vector.

Anton: For a while there's been folk wisdom, and it's actually even in the open AI cookbooks, open AI cookbooks. There's been this folk wisdom that a linear transform, which is just a matrix multiplication, is enough to transform one embedding space into another embedding space, as long as the semantics are like broadly the same.

Anton: And there was recently another paper about this where they demonstrated that like, yeah, empirically models tend to learn the same representations of the same thing. So mapping between them is a straightforward operation and that kinda means that like, that's, that's very promising for, for being able to do things like that and integrate it really well with what we're building.

Nathan: Relative representations. Is that the paper you have in mind or? I think so, yes. Yeah, that was a fascinating one. You see these kind of, I mean, obviously it's a higher dimension than projected back down to an even lower dimension for visualization, but you see these kind of shapes where you're like,

oh wow, that's very strikingly the same subject to a rotation or, you know, a dilation or whatever.

Nathan: I thought that was pretty awesome. The, the clarity that they achieved in demonstrating that I thought was pretty cool.

Anton: It was great empirical work, and honestly, I think more machine learning research needs to take that empirical approach. I think machine learning people need to take more cues for biologists and really observe what's going on.

Anton: First. There's one extra interesting thing about, you brought it up just now, you, you saw how like, oh, there, these embeddings seem to be in variant to like rotations or scaling, right? So there's a very interesting there's a very interesting potential approach here, which is obviously, you know, not obviously, but in principle you could find ways to maybe decode people's information from the vectors.

Anton: So if you're a provider of say, cloud vector storage or something, and people are sending you their data you know, it might be a risk of you being able to read that and people don't wanna know that. But most of the operations that you perform in vector space are like you mentioned in variant to translation, rotation and scaling.

Anton: But that means that you can put your, even if I know that your data came from, say, openAI's a two model, if you apply like a random transform and maybe add some ways, it's very, very, very difficult for me to recover that transform, but I can still perform the same operations that you want me to perform in that space.

Anton: So, and we haven't evaluated this in depth, it's an idea that I had fairly recently, but homomorphic computation because I can perform computations for you without really knowing what your data contains. And I think that's a really interesting thing as well. And that, that might be a whole direction on its own in a little while.

Nathan: I mean the security, I'm working on a couple projects where I'm like, boy, that actually could be a great solution to otherwise thorny problem of where exactly are we sending our client data and who is seeing in what form. So yeah. That's, that's a really interesting insight and a great connection. I mean, there's so many convergence of, so another interview with the authors blip. Computer vision model, they just came out with BLIP two and Blip two

reduced the training computation requirements by 95%. They were able to train the whole thing on a single machine in under 10 days. And the reason that they were able to do that is because they kind of created an ensemble approach where they had a pre-trained vision model and a pre-trained language model.

Nathan: And they just trained this connector model between the two, I think they said was 200 million parameters. And what it predicts are text embeddings that are derived from the images. And so then they're able to inject this representation of the image, skipping the embedding layer, just going straight to injecting the embeddings as kinda a, you know, prefix to the prompt in a, in a way that is, The thing that I thought was kind most fascinating about that is they're accessing a part of that embedding space that no text could in fact access.

Nathan: We talked quite a bit about like is there way to sort of reverse that to like what text would you had to put to get those embeddings? And they were like, yeah, there's no text that would embeds, we're just finding these, you know, kinda weird spaces, but the language model knows what to do with it. But you know, it's hard not to see the analogy between like the vision model being sort of the eyes and their model being kind of, you know, some sort of relay layer and then you get the language model is executive function of some sort.

Nathan: You kind doing a similar thing, right? Like you're connecting. I mean, is that how you sort of casually would describe it to people? Like the language model is kinda the, you know, the executive function or the sort of top level processing and then you're helping get into deep memory.

Anton: Yes. I think that's a reasonable analogy. I think another paper that's in that vein you just jog my mind, is Robotics transformer. R t one came out from Carol Houseman's group at Google Brain Robotics. And what they do is, yeah, they have like a perception subsystem that sits on top of everything and then they have this planner and the planner just basically learns from these multi-head embeddings to like output plans and actions.

Anton: It's really, really cool. And it's kinda similar to the system that you described.

Nathan: It's all coming together pretty fast.

Anton: It's, it's really an interesting time to be working. I think that like, yeah, I think that there's also just untapped veins of research here. The reason, you know, we noticed this very early on, even before we started Chroma, that the

incentives between, let's say academic ml AI research and industrial academic AI research, and I don't mean universities versus, versus like industrial research labs cause they're mostly doing similar work now.

Anton: I mean more like the sort of work that production machine learning deployments are interested in versus the sort of work that maybe pushes you towards AGI is fundamentally different. One of those big differences is in academic research, there are accepted community benchmarks and your aim as a scientist is to demonstrate the performance of your model on these accepted benchmarks, right?

Anton: But the benchmarks are static. Whereas in the real world, the data's always changing. And you know, questions of like, should I train my model? What do I train it on? Is it working better or worse? Monitoring it. Measuring it are much more nce than demonstrating performance on benchmarks. And kind of, you know, we founded Chroma with that observation in mind that there's actually a lot we can do and mine, some of these retrains of research if we just focused on them in the first place.

Nathan: Going back to nearest, you know, approximate nearest neighbors, we talked about kind of various tunings you can do, where you can kind of pre-compute some stuff depending on your workload. Where you can learn these transforms to get more and more relevant over time. Do I still have to worry if I'm doing that, that there's some sort of systematic blind spot in my system?

Nathan: Like if I'm, you know, it just to start to bridge two stable attribution. If you've applied all these techniques and I upload, you know, an image that I got out of stable diffusion, is there any, is there any way that we could start to assess like whether or not there's some sort of systematic bias there?

Nathan: Or is there, is there some part of the latent space that's like, you know, being kind of unfairly not you know, not brought to the for and the attribution?

Anton: I think so. I think so, and actually again, there's an experimental feature we're deploying into chroma very shortly, which will tell you the relevance of what's been retrieved.

Anton: So you can actually do something about it. You probably, you probably know there's like a two-step retrieval process as well. It's called HyDE. Where instead of just embedding your query, what you do is you send your query to a

language model. You let it hallucinate, whatever it wants a response, but it looks more like a response.

Anton: Then you embed that and do the query. But it's expensive. It's expensive to do two lang large language model calls. We did some napkin math not long ago, and we figured out that, for example, one company who we've been talking to, if they did it if they did it even the vanilla way, it would be hundreds of thousands of dollars per week, right?

Anton: So reducing the cost of this is actually really important. And so what we realized pretty quickly was we could make this HyDE approach more efficient if we could algorithmically figure out when the results would be relevant or not. And so we're deploying that very shortly, and it's a very general idea that I think people will also integrate into their sort of app development flow it and it again, because all this [00:32:00] stuff is geometric and nicely continuous, we give you a probability instead of just a, just a threshold dsor.

Anton: So yeah, but again, pretty much everything we're talking about, more research, more work needs to be done.

Nathan: Tell us, give us the brief overview of stable attribution and then, you know, those who wanna go deeper on that topic specifically. I Definitely recommend the show that you did with with Packy on the Not Boring podcast feed.

Nathan: But give us the, you know, the kind of short overview and then I have some questions that you didn't get to cover on, on that conversation.

Anton: Absolutely. So the short version is fairly straightforward. We saw a need and a possible technical approach to figuring out, or at least starting to figure out how images in stable diffusion training set influence particular generations that are created.

Anton: And the way that we did that is by doing various by extracting various information both about like, like sim just similar images to a given generation, the training set, but then also like figuring out what the right [00:33:00] similarity is. And the reason we took this approach is again, because there are optimal approaches, but they're all computationally intractable.

Anton: For example, like one more optimal or more principled way to do this might be to remove examples one by one until, and like retrain the model every single time until, you know, the image has gone far enough away, you say, okay,

those images influence the most. But of course, every training run of stable diffusion costs about 160 grand.

Anton: So it's computationally infeasible. We found a different approach. We think it's principled. And we applied that we leverage other, you know, various properties of how diffusion actually works. Alongside like similarity and space. We construct like a what we think is a principled similarity metric from there.

Nathan: Cool. So give us just a little bit more principles behind it at, and then then I'll get into a couple questions.

Anton: Yeah, sure. Assuming, you know, assuming the audience here knows basically about how latent fusion models work. They're trained in their training set. They have image, text pairs and they're trained to reproduce.

Anton: And it the same encoded latent as the training example through the diffusion process. They're trained by iteratively adding noise to the training examples, latent vector representation, and then training a network that reverses that noise. And what's great about it, what's actually the big breakthrough is this is very efficient, both in terms of how much data is used by dimension in the vector, but also because you can train in parallel each de-noising step.

Anton: That's the big thing. So, you know, there, there's a few parts to this. The starting point is like, okay, well given that the model's training objective is minimized, if it can exactly reproduce the latence of all its training examples and cause we know that the latent is continuous smooth by definition because that's how works you have to be, that then it's not unreasonable to say that training examples that are near the generation in latent space in some way are influential to the output of that generation.

Anton: So that's where we started from. We took a look. We found out a, a bunch of interesting stuff. We found that for example [00:35:00] and you know, there, there was a bunch of papers that demonstrated that actually stable diffusion totally can reproduce exactly or very close to examples in its training set, which people denied for a long time.

Anton: I don't why now, now they tell me like, oh you can reproduce a tiny number of examples. Like, it's not what you were saying before... but anyway, we found that noise in the training set is definitely something quite interesting here. Cause the clip representation of images and text might differ for give example.

Anton: And you can imagine why the same image might have a lot of different captions even in different languages. And then a lot of the data's just noisy. So we took that into account and we took into account. The sort of the attention mechanism of the diffusion model itself by constructing attention maps.

Anton: And attention maps are actually, like if you, you look at them in a slightly different lens. They're actually just, again, a matrix multiplication. They're a linear transformer, that vector. So when we apply the intention map, we get, you know, it's as if we're transforming the latent space and then we can do similarity search. That's like a very short overview.

Nathan: So yes, again, for more depth on that, go to the Not Boring Anton teaches Packy about AI episode six and there is quite a bit more detail there. So just kinda zooming out from the technology little bit, I naturally went online and tried it and tried it with a mix of images that I have made with diffusion.

Nathan: I know that it's not really meant to be used on natural, you know, images that I took on my phone but course I tried that as well. The Divergence-- definitely the real images, you know, like stuff on my phone and what comes back is, is bigger. And I think you have an explanation for that around kinda or why it doesn't come back with so similar images, even though I, I mentioned that they must exist, right?

Anton: sure. Here's the interesting point. The similarity is not perceptual human similarity. It's models. It's how the model perceives the difference. That's an important thing to note. It's, when I say influence, it's not necessarily the same way that an image might influence a human artist where they'll draw motifs and inspiration from, but it's more like this very mechanical, raw interpretation.

Anton: Like these vectors are the same , right? These vectors are similar. So to human, to a human perceptual system. Once those images are decoded from their vectors, they might look pretty different and you won't know what's similar about them. But we know for a fact that like ML models are kind of weird in the way that they think quote unquote, about the world.

Anton: And they might represent something quite differently to what a human would, they might focus on quite different things because they're mechanistically pursuing an objective. They're not trying to encode any meaning. And I think that's what producer surprises. I think one way to think about what you get back With inserting images that are not from stable diffusion is if stable diffusion had produced this.

Anton: So if it were, if it were possible for the model to produce this, or put another way, if you like, if the model had produced this image with a corresponding caption that was appropriate to it, here's what would have been in the training set. It's like it's a long chain of counterfactuals. It's not really meaningful.

Anton: So here's, here's the thing right in, and you can go look at this. Anyone can, if you go to GitHub and you look at the stable diffusion code, whether version one or version two as provided in the open source repo, there is a watermark function in there. That watermark is not present in most hosted versions of stable diffusion.

Anton: And we dunno why, or at least if it is present, people don't tell us what the watermark is. And we dunno why that was our plan originally to sort of reject non-stable diffusion generated images, but it's just not present. And we felt that better to release and get feedback than not. Yeah. You know, actually I think there's an interesting experience. When I used my real, you know, taken on my phone images, the re the results were kinda, [00:39:00] you know, I don't wanna use this term, but it's the only one that's coming to my mind. I know it's loaded. But it almost did feel like the pieces that were kind of collaged together to make my actual real image.

Anton: Like I tried one where it was just my kids saying goodbye to their friends at our front door. And what I got back was not like all, you know, kids at front doors, but you could see these like elements that were very resonant across the images. And I felt like, you know, there's almost something here product-wise that is kinda interesting. Like it's a sort of diverse image to image search that like picks up on these kind of notable aspects.

Anton: Have you seen? I think it's like same vibe or something like that. It's like essentially that it's like a mood board search engine engine that I think uses similar principles under the hood. I think there's a bunch of ways that you could take this.

Anton: I'm actually like, I'm curious, you know, if we do keep doing experiments in this direction, maybe we, right now, obviously we're focused on other things. I think this is a great chance to turn like generative models into something besides a one-way process. Like if you can manage to, to do these kinds of relations, you can really, you know, start to use them more as in like a search or an iterative or iterative conversational search type of approach, which I think is really interesting. But yeah, the thing that you're getting back is like,

okay, this is how the model sees your image almost. This is like what it would've tried to do if it was generating it.

Nathan: It's cool because it, it does have this sort of, it, it sort of reflects back to you like, These are the things that made your image distinctive in a way that you couldn't necessarily have articulated yourself, but like, oh yeah, I see like my kid in this picture is like turning over his shoulder and looking a certain way.

And like this other image is like quite different in all the other respects, but it has that same like pose. So there's some like feature that, that they have like very, you know, very much in common even though like the other features are, are very much not in common. I actually would recommend that people go try their real images on your search notes, not what you meant to be,

Anton: Please don't yell at me if you do that. Yeah, fyi. It's same.energy is the same.energy. Same.energy. I really like that. I think it's fun.

Nathan: Okay, so let's return to the intended use case. So you put an image that you made with stable diffusion into the product. It does. Its. Searching and you know, principled extrapolation of the search comes back with images.

Nathan: Sometimes I see a few, sometimes I see a lot. I was really left wondering like how are you deciding how many to include? Is there like a certain cutoff or are you kind of doing like a top P type of [00:42:00] thing? What do those distributions look like? Yeah, that's a really

Anton: Interesting question. It's something we've tried to instrument, but it's again, one of those things that you need a lot of data to figure out. And a lot of compute, which we're not applying to this right now, but it's interesting, you're right, sometimes you get not many matches and sometimes you get quite a lot. And I think that that depends on where you land in, in this like modified similarity space that we talked about, like you mentioned earlier with the language model where it can like generate reasonable text from places that are nowhere in its training set.

Anton: You can think of this almost as like a dial maybe from the, the model reproducing stuff to the model generalizing, right? If you, if you can kind of maybe look it up at that from that lens. And the thing is, is the stable attribution algorithm really doesn't take model generalization into account. We we're trying, we're like leaning really hard on this, like, reproducibility rather than, than generalizability.

Anton: And there's a lot of questions to explore there. And I think that they will get explored because we're gonna look for ways to continuously improve the performance of the models outside of the scaling laws. Although a lot of, you know, the labs who have the most compute, they'll keep pushing on the scaling laws, the labs who are looking to just, you know, do something interesting.

Anton: Or even, even at the product level, people are looking to do something interesting. They'll experiment with stuff like this and be like, oh, okay, I noticed that like this part of the generative space doesn't have much in it, but people are asking for a lot from here. Maybe we should like go out and.

Anton: Get stuff that, that people want from, from this part of space and then like get people to actually make images for it. It's an interesting possibility for the future.

Nathan: So when I do get those results, is it, is it fair to say, I think you said this on the, on the Packy show, that the, you know, everything influences everything in this space, right? Cause like everything's involved in gradient to say at some point. So in theory if you had the kind of fully principled answer, they would all be, you know, infinitetesimal, but probably non-zero.

Anton: Yeah, it's an interesting question. I'm not sure that that's quite right. I don't think that they would necessarily be infinitesimal. I think that different generations will draw from different parts to greater or lesser degrees because of the training objective. Further things that are further away should naturally not.

Anton: Influence the generation from a particular conditioning caption as much as things that are, you know, related to that conditioning caption. I think that, like, I think that's also a pretty interesting question. Cause one open question is in what way does generalization actually influence the output? Like we, we can't, we haven't looked at that.

Anton: I think that you could do this kind of research with a toy model and really see like, oh, you know, just have a, just have a tiny diffusion model. Make, make pictures of flowers or something and actually just, just literally just burn a bunch of compute on figuring out like, ok, this is like proportionally how the data set actually does influence things.

Anton: I think there's another really interesting approach here, which you can take with a toy model, which is basically like, compare the stable attribution approach with like the gold standard and see how good or bad we actually if we had GPU time, we would just do that. And then I'd be like, ok, yeah, everybody

was right. This sucks. Or I could be like, oh, actually know this is really great and I have numbers now. But until then it's, you know, these questions are pretty hard to answer.

Nathan: So in the process that you're running today, does it generate a score for each image that it finds like they're ranked? Right? Does that have a clear relationship to like the whole, like do those sum to one or do they sum to anything that, you know, what they sum to?

Anton: I mean, it's, it's, there's, there's some metric in space. It's, we treat it more like an ordinal thing. And then you sort of, you know, you can get a proportion out of that by you say take the top end and then you divide evenly according to according to the similarity metric. But it's not that meaningful.

Anton: It's much more of an ordinal question. You can say, okay, these, these are like closer, these are further apart.

Nathan: So then what is happening when there are no results? Like, do you also have a threshold below which you just are like, nothing is close enough to this.

Anton: There's a few filters going on. Behind the scenes we have a threshold where we just stop looking. You know, it's pointless to get things that are far away. We say that, okay, really only this can, should be regarded as influential. It's an arbitrary choice. It's a heuristic. And again, that's like, okay, well here the model is really like extrapolating and filling

Anton: the gap. So I'll, we'll leave aside for the moment, like why, you know, deep dive into, for, for, you know, sort of an economic dimension to this attribution that could, you know, when day come to exist.

Anton: Again, you talked about that in the recent show so don't want to rehash it. To what degree this is separable. Like I was just running this thought experiment where I'm like, okay, let's say I have a stable diffusion image and you know, Greg Brokowski is in my prompt and it's amazing. Like you go on these sites and you still look at the prompt and it still feels like one outta every three.

Anton: As like Greg Brokowski, you know, by name. I think I got stuck in people's heads as well early on. It's amazing how like the very early things that people discovered, they just like become gold standards. Even though it could have been anyone, it could been any choice. There's plenty of concept artists working today. It's kinda, it's, it's one of these artifacts. Internet culture.

Nathan: I guess I'm wondering like if you had something, if you're doing this now, they've mechanism and all that kinda stuff, to what degree is everything sort of meshed together and not really separable? So like a thought experiment would be, let's say you took all the stock photos and then you added all Greg Murkowski's real art. Do you have any intuition for whether I could still get Greg Rutkowski stuff out of that, or do I need like a richer, denser space?

Anton: Yeah. Look, this is exactly the kind of research that needs to get done. This is exactly the kind of research that hasn't been done to this point because we've all, we've been working towards the capabilities of these models and we need to start looking at like, okay, like controllability and engineering principles.

Anton: The way that I think about this is in the early days of aviation, we had like intuition about what makes airplanes fly, right? We had, you know, the first powered flight, the Wright brothers were wrong about how the right flyer worked. They just had intuitions about it. That's an object demonstration. It doesn't really matter if your intuition's wrong, if the thing works.

Anton: And then there's this Cambrian explosion of different types of aircraft. People were putting two, three wings on them. People were, you know, experimenting with different materials, different fuselage shapes that. Crazy fans in a tube because we, all we had to go on was intuition. We didn't have engineering principles yet, and we didn't even really have the tooling to develop those engineering principles.

Anton: That's where we're at today, in my opinion. With machine learning, we have a whole bunch of intuitions. We kind of know how things work and what's really interesting, and this happens very frequently, is like some paper will assert and you know, even with stable diffusion, even, even with like, latent diffusion models, they were like asserting, oh, this is like lagovan thermo diffusion dynamics.

Anton: And the paper came out the other day. It's, I think it's called ColdFusion or something like that, or cold diffusion, where they're like, no, it's not. Here's, here's an object demonstration that that's not true. We're in that era of just experimentation. None of this stuff is like really well understood, and we're seeing the early days of people trying to develop principles.

Anton: The scaling laws were the first thing where it's like, okay, well we're gonna empirically demonstrate that you need to increase data as much as you increase compute. Otherwise it's worthless. That'll develop further until we

probably get to a point where we can actually design. The things that we're building for the mission that we're intended to perform.

Anton: Like we design aircraft today because we've, you know, and we still don't know everything about flight. Flight is a very complex discipline, but we now have the tools and we have enough experimental sort of evidence and we have enough, you know, laws of physics that we can really like build pretty amazing aircraft. But it took a long time.

Nathan: Yeah. For what it's worth the, it sounds, you might have read it, but the David McCullough;s wright Brothers biography is pretty short and absolutely fantastic. I could not recommend that. Would you send me that? It's very striking just how the, you know, these two brothers cared about basically nothing but doing the work there was and their family, but they really did not.

Nathan: There were a lot of folks out there who were like hying themselves up as the great experts in flight and they never actually flew anything despite their, you know, supposed expertise. And these two dudes just, they put their heads down and they just did the work until they actually had the goods.

Anton: Yeah and then they, then they milked it, then they milked it through intellectual property, Laura, just as far as they could until Curtis bought them. It's, it's kind of a, it's kind of a Greek tragedy, honestly. It's interesting.

Nathan: That's not as much covered in the book, so maybe I'm missing some of the less flattering aspects of their story. I wonder if you have any, any take on the kind of technology angle. You know, I, I was thinking also as I was doing this, you go through all these prompts and you just see, you know, 4k, unreal Engine, you know, Pixar style, you know, red brand cameras, and, you know, envisioning your kind of imagined future of like people getting, you know, some sort of economic remuneration for their contribution to this.

Nathan: Do you think that like those sorts of technology layers also have a claim. Like, it's interesting. I think that's interesting. That's a really interesting point. So, so first of all, I like, for me as somebody who works in this discipline, prompt engineering kind of feels like, and I've said this sort, kind of feels like trying to pick a lock with a wet noodle.

Anton: Like there's a behavior that you, the model you have that the, that you want the model to do, but you're the only interface you're using to the model is like literally typing text into it when what you really wanna be doing is like,

have feedback mechanisms inside it. Like, you know, with control systems, like we have an industrial machinery or an aircraft.

Anton: So like what actually is adding 4k doing open question? Nobody's like, it hasn't been answered. This is more of that kind of research that I'm talking about that needs to be done. Like what I like, why does 4K matter?

Anton: It's kind of like with, with instruct GPT, why does telling it that it's the world's expert in something make it produce better answers?

Anton: So what's actually happening? And the danger is anthropomorphizing these models, it's like, oh, it's thinking better. It's, it's not a person, it's a machine. Dissect the machine, see why it's doing that. Where is it going in its latent space. All these models are, are encoding and decoding a latent space. To answer your next question, like, you know, how do you attribute rights holders to equipment?

Anton: I mean, you know, it's, that's fairly settled. You don't although, you know, could be interesting. Personally, I'm not a huge fan of adding additional copyright law or additional, like strengthening derivative work IP law. One of the reasons that we did this project was we didn't want Disney to own everything forever in the inevitable copyright backlash.

Anton: As you can imagine, like, okay, derivative work, copyright law gets strengthened, and then the next day Disney says, well, anything vaguely superhero looking is in the Marvel universe. So I'm sorry you can't draw pictures of that anymore and put them online. Sorry. That's one of the reasons we actually released this is just, you know, it's actually align incentives here.

Anton: I haven't thought about that angle. I think traditionally sort of the medium is, is eightedl. I don't think that people really have claim to images based on the equipment used to capture them. Could be different here. I dunno. I can't say that I have an opinion.

Nathan: Yeah, I don't really either. And definitely I agree. Quite settled when it comes to you photographers own their work and they, that's kinda the end of that transaction. But this does feel like it could be different. You know, the, certainly just the number of times you see that as you go through the prompts does kinda suggest like, gee, there is something to that that is, you know, highly desirable that they've created that, you know, people have a hard time getting without them.

Nathan: So I wouldn't be surprised if the you know, the red camera makers attached to something at some point

Anton: look like, I mean, I think maybe it's something similar to how photographers will share, you know, their equipment. They'll say, this is how I captured this photo. Right. They still captured the photo, but there's some credit to the equipment they use and so that other photographers can learn from them.

Anton: I dunno. Like could be, yeah.

Nathan: It's a lot of free advertising for them right now as well. So they, they might not wanna opt out. Prompts are like this first, they're like zero stage of, of like attribution. Like if you're typing gradkowski, you should probably like credit the guy somehow. The, the role, the importance, you know, the, the problems associated with noise I think been one of the concepts I've been thinking about the most since I listened to you talk to Packy.

Nathan: And you know, you said that for one thing, these massive web skill data sets, like you have a lot of image taxpayers where there is high similarity and that basically means like the text describes what is in the image and then you have a lot where, you know, it's kinda like a totally different thing that may have nothing, almost nothing to do what is actually in the image.

Nathan: And that just serves to add a ton of noise to the data set. And then you're kind of extending that to say like, that noisy portion of the data set is helping it generalize. I don't have a great intuition for that, so I'd love to understand your thinking there a little better.

Anton: Yeah. So like in, let's say, old school machine learning, even though it's been around since 2015. In, the prehistoric times, in the times of our forefathers, one thing that people noticed very early on was that neural networks were very important to overfitting. And the way to prevent overfitting was to inject noise into the models training loop. So, what you would do is, like drop out was a very common thing where you just like disconnect connections between neurons while it was training so that it didn't like end up relying on any specific connect.

Anton: Too much or, you know, all these, all these little tips and tricks and techniques, which kind of got obliterated by, by just scaling, just make them, just make the model bigger feed it more data. But at the time, this, this was like a huge issue, right? And so the way to maintain generalities so that the model, like once you evaluated on your test set still performed well was to inject noise.

Anton: Like you, you know, you'd fuzz the images you would like, flip them around, try to do like plausible things that might exist in the real world. And so that's the analogy that I draw here. Noise leads seems to lead to [00:57:00] generalization in some way. It seems to allow it to, when the text and the image are not very related, it seems to, it seems to grab something conceptually here from those images.

Anton: It almost picks like supporting structures out of them without necessarily attaching the associated meaning. That's just an intuition that I have. I can't say for sure that it's real. Another way to look at it is like you can imagine one image, many captions, it allow the model to come to the same image for many conditions.

Anton: Imagine, imagine you actually had one image, but you model one image with many, many different captions. It would always end up in that image regardless of what you put into it. Maybe that's also helping like weight, the kinds of generations that you get. Like suddenly it suddenly is able to connect one image concept with many textual concepts.

Anton: Things like that might, might be helping it. Again, this is [00:58:00] exactly the kind of research that needs to be getting done. Well, hopefully I want it to get done. I dunno if other people do. It's a target rich environment right now, that's for sure.

Nathan: If if we go, you know, an hour and I'm able to that's almost like my KPI for this show is how many times I can get my, our guests to say that that research needs to be done. I think we're at like three, so I'm doing alright today. So, just briefly going back to your, your product with chroma and the Vector database, and you had said at the top, you know, there's all these sorts of, all these different kinds of things can be represented as vectors. So we've largely talked about text, which is gonna drive this like retrieval future, where, you know, the AI's gonna have seemingly like photographic recall of all of our emails and our, you know, interactions and so on.

Anton: Even better than that, put, put that information together in a way that actually answers the, the query that you have, answers the question that you have or like, fits the context that you want it to operate in.

Nathan: So what else is, is getting embedded these days?

Anton: Oh boy. I mean, you guys have probably heard of Whisper before you have a podcast. So yes, whisper is audio embeddings.

Anton: Any, anything, basically any modality that you can stick a neural own network onto and preferably a transformer at this point onto you can get embeddings from it. So we've got audio already. We've got video. I've been speaking to a couple of people who have been working on that, and it's actually starting to get commercialized too, like video embedding.

Anton: So you can do like semantic search inside a video. You can be like, find me all the, find me all the, like, footage from this party that's actually fun to watch and it'll just do that for you.

Anton: So video, you know, and I think it'll just, I think it'll keep going. I think that we'll see over in robotics for example, like chains of actions are also in embedable. Like, oh, the person like performed this task that's, that's that can, that's embedable lots more like that. Basically all kinds of modalities will continue to get embedded and I think what will continue as well is, so we've talked about CLIP *contrastive language image pretraining*. We'll see more models trained that way.

Anton: So right now we have like Whisperer, which is [01:00:00] text to audio in the same space. And we've got clip, which is text images in the same space. And we'll have video and video and text in the same space. But then you can also think of like audio to video. You can think of images to audio that whole matrix of things. And each one of those entries in that matrix is a potential application of one kind or another which is a really exciting thing.

Nathan: Are you seeing things also like from biology starting to get embedded? I was thinking genomes and protein structures.

Anton: Yeah. Biotech I think is waking up to this biotech is a very interesting field to think about this because in biotech you kind of have to get it right. It's okay if your semantic search returns things that only kind of look like the thing that you wanted in bio. You really need the protein to be the one you were looking for. But the other part of this is in bio, most of those embedding functions that we talked about at the top of the, of the recording, Are still hand-rolled.

Anton: Very few are actually, you know, trained and modeled. And I think we've seen the success of this in these other domains that we've talked about. And I think biology is starting to wake up to this pro to this whole possibility. And so I think in the next year or two, we're gonna start seeing a lot more work in this direction.

Anton: And I've been talking to sort of biotech founders and companies and people working in this space and they're all very interested to see what, what can be done here. Certainly like, you know, things like, you know, DNA transcription has traditionally been, you know, they've tried to approach this from, from an embeds perspective before, but none of those functions were trained.

Anton: So it's like, it's, it's new and we'll have to figure it out. And what's kind of cool is they get to import all of the lessons learned elsewhere about how to like, train good embedding models and just apply them to bio and then see how well it works. It's pretty exciting. Yeah, it's kinda the great convergence again.

Nathan: You know, I've been using the phrase Kurzweil's revenge increasingly often over the last year or so. And that's probably a good bridge to kinda the last topic for today. Are we all gonna die or what's gonna happen?

Anton: Yeah that's a good bridge. That's a really good bridge because you brought up Kurzweil.

Anton: And one of my favorite things about Kurtzweil is I went on the website not long ago and it was like the 10th anniversary of the singularity is near. And that just, I dunno, that made me laugh. So look, I don't think we're gonna die. I have yet to see a convincing argument that we're gonna die. Every argument that I've seen in that direction requires some sort of system of basically incalculable and definitionally like incomprehensible power, which, okay, like if you're gonna invoke magic, you might as well just do that at the start.

Anton: I think that there are dangers. I definitely think that there are dangers. I don't think the dangers look like AI, super smart overlords pursuing fat, human unfathomable goals. If you're concerned with that, you should probably work on sun alignment. Cause the sun is giant and incomprehension has incomprehensible goals and is very dangerous to humanity but it actually exists today.

Anton: What I'm more worried about is, and this, these are like cycles in history and you see them repeatedly. We see them with the inventing of the printing press, and then radio and then television significantly destabilizing to society and often either like very, very violent period.

Anton: So the 30 years war is the one I always come back to. And the 30 years war in part was triggered by the printing press. You, you know, Martin Luther was able to spread his ideas very quickly throughout Europe because they could

be printed. And that caused, you know, players in power to like be able to attach to this movement.

Anton: And they produced schism and then parts of Europe were 50% depopulated in a matter of a few years, which is, which is incredible. It's still per capita, the 30 years war is still the worst conflict in Europe, in European history per capita. You know, and then you have the rise of totalitarianism, which was, you know, you put the radio, you put a radio in everybody's home.

Anton: Suddenly a person who wants to be a dictator doesn't have to walk physically from city to city or just convince you through text. You could hear ORs. And I think it takes time for societies to develop. Basically like when I think of as mimetic antibodies, like I think if someone heard like Hitler ranting on the radio today, they'd be like, damn, this guy's crazy.

Anton: Although, you know, we do have that. There's a few people who are still pretty persuasive. But it takes time. It takes time. And there's, there's also this bizarre thing that occurred in my lifetime, which, which still deeply confuses me. Back when the web first came out, everybody kind of knew not to trust it because people were like, oh, anyone can put anything on there.

Anton: Like, don't trust anything you read on the web, any, like, people can just lie. And that doesn't, that has never changed. Like you can just still go on the internet and lie and there's like that famous meme. But people seem to have regressed, people like seem to believe what they read online more readily now.

Anton: And it's almost because it got legitimized as an information distribution mechanism that people like lost that immunity. And the reason that I bring this up in the context of AI is because like you can imagine very concrete use cases for even something like chat, GPT where you can have like finally targeted down to the individual propaganda.

Anton: We've seen the success of like social media and doing this like ad targeting, being able to be really mobilizing for like elections in the United States and in other countries. I think at the point where you can like create text that's individually targeted and kind of like start hacking people's actual preferences to get them to vote for you or to like incite violence or any number of things.

Anton: That's, that's kind of dangerous. But it's not the ai it's the people running the ai. There's another interesting thing which I've been speaking to a few people about and I'll, I'll sort of mention in the abstract, which is like, There are

dangerous things in the world, which are still gate kept because the knowledge of how to do those dangerous things is very difficult to come by.

Anton: But if you have this reasoning machine in your pocket, you don't have to be smart anymore. You can ask the smart reasoning machine that you have to do the dangerous thing. And so that's another, like, that's another real possibility. I called it a knowledge capability. Somewhere. And I've, I've been getting I've been trying to get chat GPT to tell me how to make a neutron initiator, which is a classified component of a hydrogen bomb.

Anton: They won't do it. These things never let me get my work done. These are things, these are things that are, are more worrying and then these are just the things we can see. I think the destabilizing effects of new media technologies are always unexpected as well. And I think we're in a very early stage, so we need to kinda be vigilant about this stuff.

Anton: I think that's, that's kinda the perspective. I don't think we're gonna die, but. But I think, you know, it might be a dangerous, and, and then in the in the sort of the cursed sense of interesting times.

Nathan: Yeah. That seems to me that's kinda why we called this show the cognitive revolution. That seems to me kinda baked in at this point from my survey of the AI landscape, you, we've talked from a couple different angles about how kinda the same architectures are working for everything.

Nathan: All this progress is happening in parallel and, you know, we're talking to folks like yourself who are chipping away at kind of, you know, and I don't think, by the way, I don't think anybody really thinks right now that the core technologies, like the fundamentals are gonna stop, right. Where we're, you know, or that there's gonna be no new insights or no, you know, architectural improvements.

Nathan: I expect that too. Well, even if I kinda rule that out, I think, geez, the core stuff right now seems powerful enough to be transformative. It hasn't been productized, it hasn't been sanded, the rough edges are still there. They haven't been sanded down. It hasn't been, you know, connected to all the other things. Really new being, you know, I'm still on the wait list for God's sake.

Anton: You missed the fun part.

Nathan: I know. Well, yeah, for better or worse but these things are all gonna happen, right? And so the impact to me seems like pretty baked in. I think we're in for a wild ride that we're really not ready for.

Anton: I don't think there's ever been a time in history where someone's been like, okay guys, get ready. We're gonna press this button and everything's gonna change. But I think that actually speaks to our resilience as a species. It's it that that's still literally true. Every single time that's happened, we've come through it and, you know, as some people take the perspective that this is a unique risk, I think that the actual risks that exist rhyme pretty strongly with other times we face them. So I, you know, I put that not in general as like a, as like a statement like, oh, we'll be fine. We've always been fine before. It's like, no, I think we'll be fine because the actual risk we're facing Ry with risks we have actually faced before.

Anton: The way, the way what you put right now, I think is actually a good point. And people working in, like, broadly speaking, the, the, the realm of AI and AI safety, the sort of, the, the, the consensus is, is even if we stopped training and developing new models today, we would keep finding new uses for the models we have today. Like if you, if you see how people experiment, gpt three and 3.5, like discovering that you could ask it to think step by step makes the answers better, didn't require training in your model.

Anton: You just literally like, oh, this, this thing can do that. They call it a capabilities overhang if, if you're in safety space. So yeah, we could literally stop training right now and, and still not know everything we can do already. Yeah, there's a lot of capabilities of hanging his vest. I'm really uncertain whether or not we're all gonna die, but I definitely don't rule it out.

Nathan: And I'd like to hear your kind of response to, you know, I'll just give kind a superficial like, high level version of it, but I would say, You know, if there's like a canonical argument at this point, I would say it's probably Ajeya Cotra's Post, which is called, Without Specific Countermeasures, the Most Likely Path to Agi Likely Ends In Disaster.

Anton: I genuinely read a lot of this literature. I read Eliezer's thing, AI doom I, like I've read I think people are always surprised by how much Less Wrong I've read and how much alignment forum I've read for, given what my position actually is.

Anton: I'd be curious to read that, but you know, what, what is the argument that that is advanced there? Oh, there's a lot. So you, you know, a lot and

they're, they can be subtle, but this, the way that I like to summarize it for folks is, you know, what she describes as the most likely path to agi at this point?

Nathan: She calls human feedback on diverse tasks. So kind of a generalization of the reinforcement, learning from human feedback to more and more different kinds of tasks and you know, likely multimodal, you know, types of tasks. The challenge that she sees. That does seem to me really hard to dismiss. Not to say I'm like sure that it plays out this way, but I really can't see a great reason why I should be confident that it won't, is just simply that we are not reliable raters.

Nathan: You know, the people doing the evaluation have these flaws. And you know, we know from kind of the heuristics and biases literature, you know, the cognitive, we know that people are flawed. In fact, there's a Richard No from OpenAI has this great paper called Alignment from a Deep Learning Perspective. And he brings up a lot of these great examples.

Anton: Actually, he brings up examples where like the model actually fools the human rater cuz it looks like it's performing a task on camera, but it's actually just faking it. So yeah, like, okay, so far, so good.

Nathan: At some point, it seems like, you know, these models are getting pretty smart already, right? Like they have increasingly they're, it's like now in debate whether or not they have, you know, effective theory of mind, so on and so forth.

Nathan: It seems like it's not too much of a stretch to get to a point where the models will maximize their feedback score by not being straightforwardly, you know, truthfully, honestly doing the task. But instead start to develop a model of, well here's what reality is, but here's what this person is likely to give me a score for.

Nathan: And so then you kind of create this incentive for the thing to begin to deceive. And if you create that in a environment where we don't have the interpretability chops right now, obviously to detect if it's there, certainly with any reliability, then it seems to me like you really are playing with real fire.

Anton: There's a leap of logic be made here. So we have a model, it's performing a task that's valuable to humans where rating its performance on the task. First of all, if it's performing the task that we asked it for and we're rating it on that task, what is the difference between it doing it deceptively or correctly if we can measure the outcome, right?

Anton: And I don't mean in the sense that oh, human raters, et cetera, I just mean in the sense that like it's either doing the thing we want it to do or not. And so in the example in, in the like deceptive examples from that paper that I mentioned, well, yeah, okay, it wasn't really raising the ball, but also there was nothing riding on it, raising the ball or not, except [01:13:00] our benchmarking of this model.

Anton: That's the first part of it. In, in engineering, the real test is does the machine perform the task instead of just in a test environment, but is it actually doing what it's supposed to be doing? Right? And this is where I come from in the thing that people call seem to call alignment in AI research is called control theory everywhere else.

Anton: And these things are dynamical systems that can be understood. They have actually some really nice characteristics that allow us to understand them as dynamical systems, and we know how to build controllable dynamical systems. That's why the F 16 can fly at all is because we know how to do this. The other leap of logic that's being made here is like, okay, well it's deceptive to human evaluators.

Anton: It can hack their reward function. How does it follow from there that it's all, that we're all gonna die? What are like, what is the path to that actually happening? I've never, I've never heard a clear example of this that doesn't invoke some sort of other form of calamity that if you are worried about humans could just do that.

Anton: You don't need to worry about the ai. You could just worry about the calamity itself. I get this thing where like, oh, you know, it develops an undetectable nerve toxin that triggers once it's implanted in every human being on earth. Well, you know, ok, so why not a human with a machine that's optimizing in the same way produce the same outcome or, or you know, this thing that it builds like these nano self assemblers and there's this famous post like, Create an identical strawberry thing where it like, you know, does these nano assembler things.

Anton: It's like, okay, but like, first of all, first of all, as far as I can tell nano assemblers are science fiction, a lot of people like Drexler. I haven't actually seen a demonstration of this technology. But second of all, like, okay, like you, you don't need the AI to be worried about the nano assemblers at that point. So it's like not clear to me that an AI that okay, is like reward hacked. Some humans is any more dangerous than an out of control bulldozer. Like, we know, we know how to deal with this stuff.

Anton: We ought to think about it in those terms. I think the real danger here is anthropomorphizing the system and saying, no, it's like, it's thinking, it has goals. It doesn't have any goals. It's a, it's a computer system. It can, it can be anthropomorphized to have goals, but concretely it's doing matric multiplication and applying non-linearities to them. And that's a system, that's a, that's a non-linear. Dynamical system that's continuous and differentiable, and we know how to deal with those.

Anton: It just doesn't strike me as a credible argument. It only strikes me as a worrying argument when you start to turn these machines into people.

Nathan: Well, I worry about the whole spectrum and I get, you know, ambivalent on this in the classical sense of having strong feelings both ways. Like I love working with all the AI technology.

Nathan: I love studying it, and I'm super excited about the upside. I do you know, when you talk about like, what about just like a bad actor with an ai? That's very much I think I'm concerned. No doubt. There's a project out there right now called exactly what it's called, but it's basically text to regulatory RNA sequence, which starts to sound pretty insane.

Nathan: I asked a model one time to write some hard science fiction based on that premise, and it got me to some pretty strange places, pretty quick. Obviously we have so much time, we can only go so deep on this today.

Anton: There's one, there's one canonical example of that, right? There was the, there was that famous paper where they're like, they asked it to produce effective pharmaceuticals, and then they turned the reward function around and just started producing deadly poisons almost immediately.

Anton: Like that's, that stuff is more concerning. Ultimately, in order to kill us all, whatever does that has to obey physical law. It's, it's a large leap for me in the physical world especially, and I think this comes from my background as a roboticist in that it's really hard to do things. It's a lot easier to do things with information than it's to do anything in the physical world, regardless of how often people analogize things like, oh, DNA is just like binary encoding, or, or it's like crispr, like programming biology.

Anton: It's not really like that in the real world at all. To give another analogy, right, like anyone can make Saran gas. The problem with Saran gas is not making it, it's deploying it, it's getting into getting it into places. That's why we don't have constant saran gas attacks. Like whatever this thing is that's

supposedly gonna kill all of us has to be able to act in the physical world in a way that I just like, haven't seen any system capable of doing yet.

Nathan: I hope I never meet a robot that has the regard for me that new Bing has had for some of its early users that I can say for sure. So we're running outta time. I would love to check back in, you know, at some point down the, the line and see how both your company is developing, whether or not you've continued this, this stable attribution project.

Nathan: And we can also check in again on how how much existential confidence or nervousness we have. But a couple just real quick hitters to close us out today, one AI tools in your life. What are you using? What has changed your workflow? What do you recommend?

Anton: Yeah, chatGPT has sped up my programming work a whole bunch because whenever I work with an unfamiliar library or get a weird bug, I, I ask chat g pt sometimes it hallucinates stuff, but at least helps me look in the right direction.

Anton: I use like coding assistance. I've been using co-pilot. I recently tried cadium or pretty good, definitely sped me up as an engineer. Mostly because like I no longer have to like Google or go to stack overflow or look for documentation of stuff that I'm unfamiliar with. And what that means is it's not replacing me as a programmer, it's getting the stuff, it's getting the things that I really don't know how to do out of my way so that I can focus on the things that I'm actually really good at. It's, that's been a huge, huge productivity boost. Yeah, I think, I think that's really the main ones that I'm using today.

Nathan: Also recommend replit and their ghost writer and ghost writer chat.

Anton: Yeah. Pretty cool. We, we really like Replit.

Nathan: Quick hitter number two, hypothetical scenario, if a million people already had a Neurolink implant and it would allow you to type as quickly as you can think. In other words, you now have thought to text, would you get one?

Anton: I think, and I've thought about this, I think if we end up living in a society where somebody makes me implant a computer in my head, or it's like socially unacceptable to not implant a computer in my head, I'm going into the woods.

Anton: So that's a no. It's a strong no. Listen, I've seen, I've seen how software is written. I don't want it anywhere near my brain.

Nathan: Yeah, it's wild. I do think I kind of expect to see that day. You know, I, I used to think my grandparents who were born around 1930 and you know, my grandmother's still alive.

Nathan: She's gonna be 90 this year. I used to think that they saw way more change in their lifetimes than we ever would. And that perspective has started to change recently. And now I'm like, yeah, maybe I better be getting emotionally prepared to have a hole in my head in a little implant look,

Anton: I know what's gonna happen here. What's gonna happen here is I'm gonna have some neurodegenerative disease when I'm like 85 years old and they'll be like, install Neurolink. And it just fixes you. And I'll be like, God damnit, remember when I went on that podcast? But you'll remember it after they do the implant.

Nathan: That'll be the real tragedy. I think you're right. Alright, last one. You know, you're not so concerned about the, the doom scenarios. Give us your big picture hopes, positive hopes for, and also to the degree that you have them downside years for AI for the rest of this decade.

Anton: Yeah, I think that's a really good way of framing some things.One of the problems that we faced historically as a species is the thing that we talked about earlier where nobody tells us like, we're gonna push a button and everything's gonna change. But I think AI gives us the possibility of having tools that allow us to rapidly adapt to our situation.

Anton: And kind of the fundamental thing that humans have going for us is our adaptability. So a prosthesis, an extra tool that allows us to adapt even faster, makes it easier for us to deal with problems as they emerge. It actually could make us safer in the long run. And the way that I think about that is like right now, for example, to support, suppose that there's like some frightening thing like, like covid, right?

Anton: In Covid you needed a PhD or or an MD to understand like what to do with it, right? And there's only so many people in the world that really understand like what recombinant DNA really means. I certainly don't, I'm not a biology person, but if we had tools to like get people to the research front, like a PhD in a box where it was like exactly tailored, individualized to you and turn you into a researcher just cause you wanted to work on that stuff.

Anton: That sounds like we could be much more agile and nimble as a species to deal with anything that might come up that's like, that's a hope. The, of course, the downside is everyone kind of, every new media technology has this other property where at the start everyone believes it's gonna be for education, but it ends up getting used for pornography and propaganda.

Anton: And I think this time is no different and the propaganda part worries me and the sort of this new idea of like knowledge capability through having a reasoning machine in your pocket kind of worries me a bit. The thing that doesn't worry me is like, I don't think that, like you, you, you could probably reason your weight to some very dangerous things, but again, those very dangerous things require you to be able to act in the physical world.

Anton: So it's, we're talking about more like fairly organized actors being dangerous rather than individuals. Hard to say. But yeah, like overall, I think also, I think one thing that's gonna happen, I think one other dangerous thing which we didn't really discuss is there's a very real possibility that another AI winter will happen.

Anton: And I actually asked Sam Alton that question a little while ago. There was some sort of event. Honestly over exuberance about the capabilities and then they've failed to deliver it. That's what's happened every other time in AI winters, right? Like, you know, the previous one was like expert systems and one before that was neural networks.

Anton: Now we're back at neural networks again, that's kinda worrying. Because another, another fallow decade when we could have these incredible technologies is actually like that. That seems really bad to me. We should, we should try to like be fairly clear. Right? That said I'm still bullish because there's, like we said, there's like these massive capabilities overhang we haven't even started to explore.

Nathan: So that's all the time we have for today. Antonov, the company is Chroma the embeddings database. Thank you so much for joining us on The Cognitive Tevolution.

Anton: Thanks for having me guys.