uberknowledge

Emrah Gultekin, CEO, Chooch AI

Working with Data While Respecting Privacy

Emrah discusses the best way to store data, the privacy paradox, and cautions against overregulation.

00:55 Chooch AI’s mission is to emulate human intuition.

06:00 There is a privacy paradox and the best way to address it is to score the benefits and disadvantages of having your privacy revealed.

08:19 The Western World is very privacy driven and it is creating tension in the tech world

13:18 Don’t keep raw data. Use neural networks to hash it.

14:13 Governments do need to regulate, but they should be cautious. Overregulation stunts innovation.

18:34 People overemphasize data. What is actually important is how your models and platforms interact.

19:45 Ashwin Krishnan: if you could create a closed-loop pipeline of data collection, model training, and prediction, you would be able to discard the data that isn't adding value.

21:56 Use public data, don’t collect private data to train your models. That will solve the privacy problem.

Ashwin Krishnan: [00:00:34] Welcome to another episode of the UberKnowledge podcast. With me today, I have Emrah Gultekin, who is the cofounder and CEO of, hold your breath, Chooch. Can't forget that name. So Emrah, happy to have you on the podcast today. Why don't we talk about Chooch?

Emrah Gultekin: [00:00:55] Yeah. Glad to be here. Chooch AI is a visual recognition system. It's an AI training platform for visual recognition. What we do is we train expert artists on specific fields and deploy them in companies' systems. So whether it's their enterprise apps or any type of public app that they have, we train those expert eyes and put them into their system. We're trying to copy human intuition, who are experts on certain fields and have them deployed to do certain work in visual recognition.

Ashwin Krishnan: [00:01:38] If you take a step back, like we were discussing before the podcast, we live in a distracted world, where too much information is probably an understatement, it's an avalanche of information. You're trying to replicate the human mind visually. Now from a use-case perspective, we're at this conference today and tomorrow, the IoT Blockchain AI conference, and there are so many people, so many things we're trying to assimilate at the same time, and it's a challenge. So if I were to take that use case of people attending conferences whether it's this, whether it's AWS Reinvent going on in Vegas right now … There was a person I was talking to earlier and they said, "I have an increasing sense of inadequacy with every conference I attend because I feel it reinforces how far, far behind I am." So, if you take that context and say from a human perspective there is only so much memory and cognition that a human being can have — you have all of this source of information being dumped at you — how do you see use cases where something like Chooch AI can help solve that?

Emrah Gultekin: [00:03:01] Yeah. So we're all being overwhelmed with a lot of information. I think a lot of people confuse information with knowledge, and information is not knowledge. Knowledge is basically the extraction of that, the extraction of those features. Basically, what we do at Chooch is we try to make things simple as well. Whether that's a simple understanding of authentication or whether it has to do with a certain piece of visual content that needs to be seen, we try and simplify that into very small tags and small stories. So summarization, I think, is very important here. Visual data, in video especially, images as well, clustering those and making a summary of that, so that people can make decisions based on it. It's the human mind. We're highly intelligent beings, but there's only so much information we can take in, and the best way to do that is to summarize.

"A lot of people confuse information with knowledge, and information is not knowledge."

[00:04:04] So at Chooch, we're trying to do the same thing, summarizing visual data for companies and for consumers as well. So that's one step that we're trying to do and make people's lives easier in order to make decisions easier for them or supporting their decisions that they've already made. So those are the things that we're trying to do. I mean, I think people are overloaded with this information, but we need to have a way to make predictions which are very, very simple for people to understand.

Ashwin Krishnan: [00:04:38] Got it. So let's put on our ethics and privacy hat for a minute. And your videos, by the way, were compelling. Just knowing that there is a wealth of summarized information that I can go back and retrieve at will is, personally for me, a big game changer. On the other hand, targets that I'm looking at, visual targets for instance, where does privacy come into play? I used this in my talk earlier today, if you have an Alexa at home and your neighbor stops by and the conversation gets recorded, at that point you are a data controller. You have to disclose to your neighbor that the conversation may be recorded. So in some sense, the consumer is now a producer because you have a data collection agent at home. What you're trying to do obviously has enormous value, I can see, but again if you were to put on your privacy hat, how would that manifest itself into a symmetrical information exchange with targets versus an asymmetrical exchange?

Emrah Gultekin: [00:06:00] So 90 percent of what we're doing is we're trying to collect public data, it's already public. So we're trying to stay away from the privacy, some privacy issues around that. But that doesn't mean that you completely avoid it. I think the privacy paradox we have today, I call it a privacy paradox because at the same time you want better, you want...

Ashwin Krishnan: [00:06:27] Service, better targets.

Emrah Gultekin: [00:06:29] You want to be unknown.

Ashwin Krishnan: [00:06:34] I want customized service, but I don't want to give you anything!

Emrah Gultekin: [00:06:36] Exactly. So there's a huge paradox. I think the way we should look at this is what are the benefits and what are the disadvantages of having your privacy being revealed. Whether these companies or whoever the organization is, even peers are using it for your benefit or are they using it to do something else? I think that's one way to look at this and say OK well how do we how do we score that? What's the score of the benefit and disadvantages and the potential harm that may come or the potential harm that's actually come? So I think we need to find a way to score that. To just to make sure that people understand that my consent is to give you all this information, and GDPR and all that, but what are we really giving consent to? I don't think the average person understands that.

"What are we really giving consent to? I don't think the average person understands that."

[00:07:37] The only way we can overcome that is maybe a tool to regulate the benefits and some of the harm that might be caused by this and to be able to legally enforce that moving into the future. It's not an easy task. I think privacy is really important for people. We have to take care of people. We have to take care of their privacy. But we have to also understand that the reason we're doing this is to increase productivity and increase efficiency for that person or that group. So that dichotomy, I think we're going to be dealing with for decades. We'll see how that develops. In the western world, we're very privacy driven and that perhaps is creating some of the tensions we see in the tech world today. For example, in other countries, I won't name them, but in some Asian countries ...

Ashwin Krishnan: [00:08:37] I can talk about it, like India.

Emrah Gultekin: [00:08:41] It's like no one cares about privacy. They do care but not enough to do something.

Ashwin Krishnan: [00:08:46] It's not in their top ten. Yeah.

Emrah Gultekin: [00:08:49] And you'll have backlashes on that in those types of communities. You'll see negative repercussions of that. I think it's important to find balance here. I think people who are in this industry, who are leaders and practitioners in this industry have to be thinking about this in more detail for sure.

Ashwin Krishnan: [00:09:10] That's absolutely spot on. Again back to my talk earlier today, which as business owners I gave this example of a tabletop exercise that the security industry has been dealing with for a long time. Which is really, as the name suggests, you simulate what if half of my customers' PII data shows up on the internet. What would we do — as legal, as marketing, as the CEO, as the board of directors — what would we do? It's a hard exercise because it puts into question the fundamental nature of what the company stands for. Disclose and not disclose, whether you do forensics first, if it takes six months, hope and pray that there's not a New York Times exposé.

[00:10:01] But to your point, I think the two things that you mentioned earlier and trying to get my head around from a business perspective is, we've talked about customer persona, and I have not seen too much of the tabletop exercise when it comes to privacy coupled with customer persona. If the breach were to happen to, let's say, half of our Asian customers, what would our reaction be? If the breach were to happen, let's say, in Northern Europe what would the reaction be? And while it seems pretty obvious when we had this conversation what should happen, in the larger scheme of things, where you've collected a lot of data because you can, you're sitting on it, algorithms haven't been developed, VC funding is up in the air, there are lots of other challenges. So how do you elevate this topic of the impact of how privacy can not only be something that you keep the regulators off your back but, more importantly, actually help you get closer to your customers?

[00:11:15] I think that while it might take decades for this privacy paradox to unfold, how does a business set the wheels in motion? Like we were talking about Apple as a company with Tim Cook at the top, they've taken a very aggressive stance. Now there are skeptics saying it's because they don't have an ad-based model like Google has or Facebook has. Regardless, Tim Cook standing up and talking about it sends a message to everybody inside of Apple. We take customer data, customer privacy very seriously. In your experience right now given what you are seeing, how does a cultural shift happen in other industries, not even tech industries, whether it's healthcare, whether it's oil and gas, where they're barely coming up to speed with tech?

Emrah Gultekin: [00:12:04] Yeah, this is a major issue, I think. Every vertical has its own regulations, healthcare industry has HIPPA.

Ashwin Krishnan: [00:12:12] Yeah.

Emrah Gultekin: [00:12:13] It's very highly regulated. I think it's good for that. It's highly regulated, but also you get this dichotomy where you can't really create the customer persona in that type of environment.

Ashwin Krishnan: [00:12:27] Or even in the case of HIPPA where in order for you to get a blood test, you're signing two forms.

Emrah Gultekin: [00:12:38] Yes.

Ashwin Krishnan: [00:12:38] You have no idea what you're signing. All you know is the nurse can't draw blood until you sign that. Now that's not really engendering trust. It's just like, let's get on with our job, not knowing that your blood test results could be part of a lab survey or something on the back end. So even with HIPPA, I mean I look at it and say it's a regulatory environment which is really to prevent litigation from happening. So you don't get a class action lawsuit. It's less about ...

Emrah Gultekin: [00:13:08] Providing structure. So that's in a lot of regulation. It has nothing to do with actual technology, but you can see where it could go with other verticals as well.

Ashwin Krishnan: [00:13:19] Exactly!

Emrah Gultekin: [00:13:19] You'll be signing off on all kinds of documents if you're buying something from Walmart. I mean, this is crazy. So what you want to do is, you need to keep the data private. One way to do that is to use neural networks to do it actually. So instead of keeping raw data, you could be hashing these. It's very, very difficult for someone from the outside to decipher because you turn them into numbers. So that's one thing that we do with clients. Whenever we train a perception that the raw data is all destroyed. It's all in a perception, it's a trained neural network, very difficult to decipher that. We're not encrypting it. That's another step that you could take to safeguard this.

> "Instead of keeping raw data, you could be hashing these."

[00:14:13] But in terms of different verticals taking different actions, I think we'll see more regulations coming out of governments. The US has just appointed a new security commission inside the House of Representatives, so they're going to write a report within 180 days to see how do we create a policy surrounding security and privacy. Well you don't want to overregulate it either. If you overregulate it, you're stemming your development.

Ashwin Krishnan: [00:14:47] Or people will find ways to get around it.

Emrah Gultekin: [00:14:49] Yes, I still believe in capitalistic, invisible-hand type of regulation. The market should regulate itself, normally, unless there's a huge market failure. In this case, we don't know where it's going to go, but we do need some regulation. It shouldn't be overregulated so that it stems the growth of these companies and these verticals. And also there is competition, there's global competition on this. So if a country or community overregulates, somebody else will go way past you, and won't have your historical advantages that you used to.

Ashwin Krishnan: [00:15:36] Yes. That's an interesting point because there's the demographics of your target customer but also demographics of your competition. AI is a great example, where the laws in China, the investment and privacy laws, as an example, are pretty low compared to Western Europe or the U.S.

"I still believe in capitalistic, invisible-hand type of regulation."

[00:16:00] You mentioned taking not the raw data but the neural data and being able to actually make it difficult to get the true customer data that was originally collected. But from a sophistication perspective, and I've heard this phrase FOMO all the time, which is fear of missing out, and because you can put sensors everywhere, the cost of sensors is low, and if you have a customer base, you can instrument the heck out of it. Everybody and anybody is collecting data. And so there are two sets of arguments. The first argument from the people who say, "OK, we collected the data, but 95 percent of the data is useless and only 5 percent is something we are going to act on." So therefore, if I don't collect data, and if I minimize data right now, I'm going to be left with nothing and back to your competition problem which is collecting data like crazy.

[00:17:10] On the other hand, it's not about collecting raw data, it's really getting to a point where you cannot reengineer it back to the source. That's a level of sophistication which is clearly interesting. But in terms of security, you mentioned encryption, but is there also an awareness issue of saying because data is "free", because we collect it so ubiquitously, people tend to, or at least the collectors tend to, not treat this with value. Anything which is scarce, obviously has more value, but this is so abundant. It's there everywhere. The point being, data minimization as a way to protect. You talked about the ability to derive the data versus actually raw data. Other than encryption, are there other mechanisms that enterprises can take, who are not up to speed right now with what percentage of data is going to be valuable eventually. Yet, just having that the base of saying this is the minimum we need to do to secure the data, even if it takes years for us to ever figure out which data is valuable?

Emrah Gultekin: [00:18:34] Great question. So I think there is a simple answer to that. I think people overemphasize data for that reason. It's because it's, "let's collect it and then we'll see what we do with it." And that overemphasis actually is a hindrance on privacy as well. What's more important in AI is not the data, and that's contrary to what everyone is saying, but we know, as practitioners, what's important is how your models actually interact with each other and how your different AI platforms talk to each other. So if we focus on that, then we could stabilize the part of the data that's important and just extract that, instead of extracting everything.

[00:19:24] You know, you have the three states of AI. First state is data, creating a data set or data sets, and then you train the model, and then you have predictions. So there's these three areas and everyone's focused on the data set, and that's really not, you know, it's only a piece of it.

Ashwin Krishnan: [00:19:44] Maybe it's because it's the easiest.

"We know, as practitioners, what's important is how your models actually interact with each other."

Emrah Gultekin: [00:19:45] Well, it's more obvious. It's really obvious. Models and how the models work, and regressions and predictions, people don't really get that. They think it's some type of magic, which it's not. But we focus on how our models talk to each other and how the different platforms of visual recognition, even, you have like six or seven different types of object detection, image classification, text detection, you have facial time series on videos, and so forth and so on. These are all different sciences, actually. And inside of those, you can have hundreds of different models working, and data is only one piece. If you can get them to work together and trigger each other and validate each other, you don't need as much data. You're only using 5 percent of the data you collect anyway.

[00:20:46] And so I think these are the things that we need to focus on in order to overcome some of the privacy issues that we've been discussing.

Ashwin Krishnan: [00:20:53] I think that's a great takeaway. To summarize, if you could create a closed-loop pipeline of data collection, model training, and prediction, you would be able to discard the data that isn't adding value.

Emrah Gultekin: [00:21:07] Exactly.

Ashwin Krishnan: [00:21:07] One last thing before we close the podcast. There's this whole emerging data-brokering industry. And back to what we just spoke about, if you are in that pipeline and you don't have the ability or don't have the potential to ever train a model or come up with any predictions, but you have this enormous dataset because you're an ISP or other in-transit provider, if you will. How does that data-brokering mindset now overlap with privacy and security?

Emrah Gultekin: [00:21:56] It's kind of nebulous right now. That's the issue. I think if it's public data, and there is a lot of public data out there, I would recommend people and companies focus on public data instead of private data because there's a lot you could do with public data that could offset some of the issues that you might have with private data. We see this a lot. We want to train a perception, we want to train a model, we have all the data in the world.

Companies think they have all the data, but they don't because it's not structured. It's all over the place. You wait. You don't use your data. We'll use our data, we'll use public data to train, and then you get rid of this privacy. There's a lot of public data out there. If you take the hundred percent of data you collected and only use five percent of it, you can find a lot of that in the public arena. And once it's public, it's easier to use. I think we need to focus on that.

"I would recommend people and companies focus on public data."

Ashwin Krishnan: [00:23:05] Very cool, I think this has been extremely educational for me and hopefully for the listeners too. Thank you for your time. I hope you have a great two days at the conference.

Emrah Gultekin: [00:23:13] Thank you very much. I really appreciate it.

Ashwin Krishnan: [00:23:16] Thanks.