Transcript Episode 9 Dr. Khaled El Emam

Ron Kruzeniski:

I am very pleased today to be talking to Dr. Khaled El Emam. He is well qualified to talk about synthetic data. He has his PhD in engineering from King's College London. He's currently the research chair of the School of Epidemiology, University of Ottawa. He's also a senior scientist with the Children's Hospital Eastern Ontario. Previously he's worked for the National Research Council. He's also a senior vice president of Replica Analytics. And he has been involved in five other companies that have been involved in data management or data analytics.

I think that hopefully proves to our audience listening that he is an expert in the area of data analytics and data management, and more particularly, as we'll get into it, synthetic data. Thank you, Khaled, for agreeing to do this podcast.

Dr. Khaled El Emam:

Thank you very much for inviting me. Looking forward to it.

Ron Kruzeniski:

So, we're going to talk about the study that you've just completed. But before we do, for our listeners, can you give us kind of a working definition of synthetic data. And reading in your report, I noticed you talked about a continuum, starting with anonymous data to the other side of personal data, and you slotted synthetic data into that continuum. And that really helped me kind of understand where it's at. So what's synthetic data?

Dr. Khaled El Emam:

Let me start off by just adding a bit more detail on the report because I think that would be helpful context. So, we just delivered a report to the Office of the Federal Privacy Commissioner as part their Contributions Program. And the objective of that report was to analyze how synthetic data should be regulated in Canada.

And now let me answer your question about synthetic data. It's not a new technique for creating de-identified or anonymized data. It's actually been around for a few decades. But it's picked up quite a bit over the last few years. And it's essentially the creation of fake data. So you start off with a real data set and you build AI models to learn the patterns in this real data, and then you use the AI models to generate new data. And you have seen synthetic data probably in the past in the form of the deepfakes. So you've probably seen fake images of people that look very realistic or you've seen fake videos where people are doing things that they never did, because the videos were generated by taking someone's video and replacing their face with someone else's face, but it looks very realistic. So these are also forms of synthetic data, except they are images or videos rather than actual data sets, like let's say numbers or tables in an Excel spreadsheet, and you can also create synthetic data of that form.

It's a type of anonymized data, but it's more than just anonymized data. Because traditionally when we think about anonymized data, we think about identifiability. And identifiability is whether a record can be mapped or assigned to a real person. So I can say record number five in my data set belongs to Ron. So then I've re-identified that record. With synthetic data, the records are not of real people. It's a generated record. It doesn't map directly to a real person. So the identifiability risks of synthetic data tend to be quite low. If done properly, they tend to be quite low. And this is one of the main reasons why this approach has picked up recently in that it provides a strong protection against identifiability in

practice. It's not a panacea and it doesn't solve all privacy problems, but it does have some advantages compared to traditional methods for creating anonymized or non-identifiable data sets.

Ron Kruzeniski:

Now, you said there, "decreases the chance of determining identity." I guess there's no such thing as a guarantee of preventing a determination of identity and it's a matter of just reducing the risks?

Dr. Khaled El Emam:

That's correct. There's always some risk of re-identifying a record or assigning an identity to a record in a data set. It's never going to be zero. Even if you randomly assign an identity to a record, you still have a non-zero probability of that being correct. So it's never going to be zero. The question is, is it low enough? And there are precedents for deciding what is low enough. What we as a society have determined to be acceptably low risk.

And certainly, in the healthcare space, there have been some strong precedents, for example, from Health Canada and European Medicines Agency and health departments in the US. And in the census, for census data, the national statistical agencies around the world have set quite strong precedents as well in terms of what they deem to be acceptable risks.

So, there are quite good precedents from reputable organizations around the world. They tend to copy from each other, so there's a good amount of consistency as well across these organizations in terms of defining in a quantitative way what is deemed to be acceptable risk. But it's not zero. If you want zero risk, then you're going to be able... All data is personal information if you want zero risk.

Ron Kruzeniski:

Right. And you touched on my first question, but what would you say the purpose of the study was or what was it all about? A little later we'll get to the conclusions. But what did you set out to examine and report on?

Dr. Khaled El Emam:

We had three main objectives. The first one was to present a unified model for evaluating privacy risks in synthetic data. So we presented the different techniques that have been proposed to assess privacy risks in synthetic data, and we summarized them and we put together a unified model that combines them all into a single framework for thinking about the privacy risks in synthetic data. So that was the first objective.

The second objective was to summarize how synthetic data could be regulated or would be regulated under existing privacy laws in Canada. And we're really looking at the federal privacy laws, so we're looking at PIPEDA and we're looking at CPPA as the main laws that were analyzed, and determining how synthetic data would be treated under those. We also compared them to the GDPR, to HIPAA in the US, and to the CCPA in California. But our primary purpose was to look at this from a Canadian perspective. And so we have some conclusions there in terms of how synthetic data would be treated and what seems reasonable and what seems like creating disincentives for using privacy enhancing technologies in general, or things that have sufficient uncertainty around them that they would in practice limit the uses of privacy enhancing technologies.

And the third component was a series of interviews with privacy regulators across Canada to get their perspectives on how they think synthetic data should be regulated, taking into account existing statutes

in the country, but also how they think it should be regulated moving forward if there was a chance to make changes to some of these laws or regulations.

Ron Kruzeniski:

I know you talked to me. And I'm just curious, were you able to talk to all the Commissioners across the country? Did you talk to most of them?

Dr. Khaled El Emam:

I was able to speak to 13 out of the 14, or their staff. In some cases they had some staff who were specifically focused on this topic or who had specific expertise on this topic, so they were the ones who participated in the interview. But I managed to speak to 13 out of the 14 offices.

Ron Kruzeniski:

And did you pick up any common theme from the Commissioners? Are we all thinking the same way, or like in Canada, are we all thinking differently?

Dr. Khaled El Emam:

Thinking the same way would be too easy. I think there was consensus. I mean, there was definitely consistency. But there were also differences. I don't think there was a unanimity on all the topics.

One topic where there was unanimity is the need for codes of practice or some mechanism that would allow regulators to have some assurance that public and private sector organizations are implementing good practices, whether it's for debt and vacation, whether it's for the generation of synthetic data, there's a need because many of the offices don't have the resources to check what everyone is doing. And even when there are complaints, it's sometimes challenging to understand or figure out what organizations are doing. So having codes of practice that organizations can follow, that are approved by some recognized authority, and where there's an enforcement mechanism where it's possible to check either through third parties or through their regulators themselves then being able to audit the practice of organizations against those codes of practice. Setting those was a consistent theme as something good that we can have that is available for all the jurisdictions across the country.

But there were other items, of course, where there was less agreement. I think there was consensus but not unanimous agreement on them.

Ron Kruzeniski:

So, I know when I was thinking about talking to you, my thoughts were I focused on my province. And as you look at this, do you see it... I saw it as more the federal Commissioners having a key role on this. Do you see it that way or do you see a role for provincial Commissioners?

Dr. Khaled El Emam:

Well, I think all Commissioners encounter similar types of issues. Some Commissioners have more resources to investigate and examine and research these topics more than others. So they will have more expertise in-house to address those issues. But I think in practice, the same issues come up across the board.

Ron Kruzeniski:

So, I know you were targeting March 31st for the report to be out. I'm assuming probably by the time we post this podcast that your report will be public. Is that the plan?

Dr. Khaled El Emam:

Yes, that's the plan. There's a process for getting feedback from the Federal Privacy Commissioner, but usually this is relatively rapid. So I'm hopeful that quite soon we'll be able to make the report public.

Ron Kruzeniski:

And anyone who would be interested probably could find it on the Federal Privacy Commissioner's website or another spot where they can locate it.

Dr. Khaled El Emam:

The Federal Privacy Commissioner will usually put a link on their website to the reports that they have funded in the section on their Contributions Program. I don't know when that will happen. It may not happen right away. But the report will be on our lab's website, which is at ehealthinformation.ca.

Ron Kruzeniski:

Ah, okay. And I guess the most important part, what conclusions did you reach as a result of the study and a few of the... I think you made 10 recommendations. But what were a couple of the really key recommendations that you're making in the report?

Dr. Khaled El Emam:

I think at a general level... I mean, let me preface this by saying that achieving complete consistency across the country would be the ideal in terms of how technologies like synthetic data and other privacy enhancing technologies are regulated. However, I think in general, certain things would be very helpful.

One is the principle of reducing uncertainty. So there's some basic questions around the use of data for secondary purposes, such as whether consent is required to create non-identifiable data sets and whether non-identifiable data sets should be regulated. And if they should be regulated, what extent of regulation should that look like? What are the obligations? And so having clear answers to those questions is helpful, whatever the answers are. I mean, I know that different Commissioners and Ombudsmen and women have different perspectives on this. But I think certainty is important because certainty allows organizations to make decisions and to make investments as well in technology about what will work and what will not work.

And it also leaves less for interpretation, which I think is another factor that became apparent was when there's ambiguity, there's room for interpretation, and the interpretation may not be consistent over time as different regulators take on that position. And if organizations are going to make long-term investments, they need certainty over the long-term. So, certainty is quite important, and it can take different forms in terms of regulations, in terms of codes of practice, and so on.

The other one I think is creating incentives. So again, putting the appropriate incentives, I mean, there's an opportunity here, at least at the federal level with privacy law reform to put something in place that's built on what we have learned over the last decade or so. And putting in place the right incentives for organizations to protect data, to use better technologies, and to use those data sets, in our case, the identified data or non-identifiable data sets, responsibly, creating the incentives for that is important

and removing disincentives is important. And in some cases those incentives are not there or potentially there may be strong disincentives for that.

So I think those two factors, incentives and uncertainty, reducing uncertainty, are all-encompassing principles that will be very helpful to enable the use and adoption of technologies like synthetic data generation.

Ron Kruzeniski:

And sometimes one report or one study results in thinking that other studies should occur. Do you see the need for any future studies, maybe a study in proposing legislation? But generally do you see the need for any other studies as a result of this report?

Dr. Khaled El Emam:

Let me frame it slightly differently, if you don't mind.

Ron Kruzeniski:

Sure.

Dr. Khaled El Emam:

One of the challenges we have... My work is in AI and the applications of AI and machine learning in healthcare in general. And Canada has invested significantly in research around AI and has become a real powerhouse globally in AI research. And report after report has identified that we are not able to convert that research into products. Into technologies that are actually used, domestically or globally, or being able to generate IP that can be commercialized. So we can do the science, but we can't do the innovation as well.

One of the factors that hinders the ability to build AI models is access to data. Because AI is all about data. You need access to health data. You need access to large amounts of health data to be able to build these models, train them, and validate them and put them through the regulatory process for approval, et cetera.

So access to data is fundamental to a successful AI industry. And I think that that pattern repeats itself at other sectors, but let me just focus on the medical sector. If it's hard to get access to data and if we're not able to ensure that our rules and regulations enable responsible access to data so that we can convert that science into innovation, then we are giving up I think a very important advantage. So to answer your question, I don't think we need more reports. I think we need to solve the problem. Many reports have been written on this issue, and I think we have a good understanding of where the problems are and what direction we need to take to address those problems. I think the time now is to do something about it.

Ron Kruzeniski:

So, I wonder if this is coming at this another way. One thing that we Commissioners hear from time to time from researchers is, "There's too many barriers or difficulties to getting access to the data." And we sort of react to that and say, "No, it may be a cultural issue, but the privacy laws are not that complex."

But do you see that if we could get past and have a lot more availability of synthetic data, that we could really satisfy those researchers that want to go further in terms of their studies and their projects? Do you know what I'm getting at there?

Dr. Khaled El Emam:

Yeah. I mean, I think that would help significantly in moving things forward. Absolutely. And some organizations like academic medical centres in Canada have started looking at synthetic data as a way to enable access to data responsibly. But I would just add one more thing. It's not just researchers, it's also companies.

Ron Kruzeniski:

Yes.

Dr. Khaled El Emam:

I mean, we need to convert that research into products. We need to generate IP. We need to commercialize that research. I think that's very important as well, because that's really where the research starts to have an impact on society, especially when we're talking about AI-type research and medical AI-type research. So we need to make these rules work for academia, for the public sector, as well as for the private sector.

Ron Kruzeniski:

So, what I hear you saying is less study and a lot more action, that's the future in this particular area.

Dr. Khaled El Emam:

Yes. We know where the problems are. We understand what the remedies would be. I think we need to start solving the problem and putting in place the... While we have the opportunity. I think we have an opportunity now. Put in place a regime that would allow responsible uses of data and allow us to take advantage of technologies like synthetic data generation to make this kind of data available. As a society, we have a lot of problems, and I think we need to be proactive in using the data that we have to try to solve some of these problems.

Ron Kruzeniski:

So, you say, "while we have the opportunity." Do you mean by that that other countries will just surpass us? When we have an opportunity and if we miss it, is it going to be competition elsewhere that just takes over or innovation elsewhere?

Dr. Khaled El Emam:

Absolutely. We have the opportunity because we're going through a period of privacy law reform. So that's one factor.

Ron Kruzeniski:

Oh, okay.

Dr. Khaled El Emam:

But you're absolutely right in that competition is also something that's very important to consider. We don't live in isolation of everything else, and other jurisdictions are moving forward with developing mechanisms to enable responsible access to data and trying to address those problems. And like I said, we're, again, focusing on the AI component, but this applies to other technologies or approaches. But

for AI, we are really good at the science as a country. We need to be really good at the innovation as well.

Ron Kruzeniski:

Innovation part.

Dr. Khaled El Emam:

Yeah.

Ron Kruzeniski:

Well, thank you very much, Khaled, for doing this. And I hope listeners develop sufficient interest to check the Federal Privacy Commissioner's website in a week or two or three and there should be a link there to get right to this report. So thank you very much for doing this today.

Dr. Khaled El Emam:

Thank you for the opportunity. We appreciate it.