accenture

THE FEDERAL CATALYST

AI's Synthetic Data Paradox (episode 4)

Podcast Transcript

Announcer: Welcome to the Federal Catalyst with Accenture Federal Services, the podcast series addressing critical management and technology issues impacting federal leaders. Each episode goes behind the scenes with our experts and others to discuss the latest research, innovations and breakthroughs shaping how federal agencies achieve their mission.

Kyle Michl: Hello, and welcome to the federal catalyst with Accenture Federal Services. I'm Kyle Michl, chief innovation officer. I'm also one of the lead authors and executive sponsors of the extension of federal technology vision, which is our annual look at the biggest technology trends poised to impact government over the next three to five years. The topic of today's discussion. Is trend 3 be unreal? In this report, we. Look at the future of AI and the paradox that synthetic data and synthetic images present. Synthetic data allows us to do more with AI. Which is true in both the private and public sectors over synthetic data also is enabling deep fakes and other forms of disinformation. This raises the question of ethics and of what steps can. Be taken to limit their impact on public trust. For today's discussion, I'm joined by two of the report's co-authors. Mark Bosch is a PhD with our implied intelligence group focused on advanced computer vision applications. Mark has worked across a number of products, which includes IARPA, where he worked on a project to improve the fidelity of satellite imagery. Shauna Revay is also a PhD, leaving our Machine Learning Center of Excellence with a focus on the national security community. May also know her as the editor of our baseline Machine Learning newsletter. Welcome Shannon, and welcome Mark.

Shauna Revay: Hi, Kyle. Thanks for having us today.

Marc Bosch Ruiz: Hey Kyle, great to be here.

Kyle Michl: Well, I'm excited to jump into it. Thanks for joining me. So, you know, we've used the term synthetic data a little bit, but let's drill down, you know, maybe a little bit. Technically, what are we talking about when we're defining synthetic data?

Shauna Revay: When we're thinking about synthetic data, it can really come in a lot of different forms, whether it be tabular data, text, images, video, all of that. We're basically thinking about anything that wasn't organically produced or collected, so instead it might have been generated by an algorithm which is attempting to replicate or mimic. Patterns and phenomena that we're seeing are experiencing in the real world. So even broader, I think some of the rendering tools and creative tools that are out there these days. I know the app Snapchat.

You know you can put filters on your face, you can add dog ears or make your face into a Bunny, and some of that in the realistic Ness of it has been pretty impressive.

Kyle Michl: Yes, Neptuna is a favorite in my house, and the dog ears are fairly popular as well. That's a good example. So I'm just curious, what's driving these real synthetic scenarios where there are changes in the last 1-2 years that are enabling this? Is it AI? What's really driving the ability to create these real acting or feeling synthetic environments?

Marc Bosch Ruiz: I think to me it's a combination of several factors. Obviously AI has had a pace of innovation that I think it has surprised many. I cannot necessarily think of a of an inflection point in the AI world that has happened in the last one or two years. It just I think it's been a steady but high pace of improvement. Overall over the last five to 10 years, but I think it is also, the reality is, is that it's a field that has gained a lot of popularity and not just from the scientific community, but also so our like creators and others that that are starting to find. Easier access to AI that that has fueled that, that sort of growth in in, in in popularity on AI and you're starting to see more and more any use cases and more and more examples of how to use this, this technology, how to create different types of synthetic data that it's really like. Fueling this sort of perceived acceleration of promise.

Kyle Michl: And would it be fair to say that clouds had an impact there and maybe some of the abstracted tooling that's available from the cloud service providers?

Marc Bosch Ruiz: Certainly, I think, like, we often think about AI. We sort of gravitate towards the software component of it. But I think we get we cannot forget the hardware piece of it and certainly cloud is one of those instances of Harvard progress like the ability of now having access to high end computer engines like GPUs and massive amounts of the build has certainly also influenced in in the ability of progressing and evolving the AI technology to the to the point that we're right now, right. So it's a combination of both the software and the hardware where cloud has definitely been a big factor, a big agent there that has contributed to that overall progress.

Kyle Michl: Got it. So the technology is getting better, more accessible, more commoditized, and we can define and use and build synthetic data in a way that's looking and feeling very real. So with all this advancement, what problems can synthetic data solve?

Shauna Revay: From the perspective of an ML practitioner, I think one of the main benefits is that the hope is that the synthetic data will allow us to develop higher performing models. So ideally, when we're training a machine learning model, we want data that's representative of all the different types of classes and examples that we might find in the real world. But a lot of times when we are looking at datasets that's not the case. Whether it's because of limitations on the data collection process or the access that we have today to begin with, it's just not as representative as we would like. So synthetic data allows us to either augment existing datasets or replace it entirely with synthetic data and hopefully help model performance. And I think another side effect of that is that it has a potential to decrease bias. So a lot of times bias in these models occurs because of the imbalance in the original data sets that we had access to, and so we're able to augment those or inject synthetic data into there to help relieve some of that imbalance. We might be able to improve bias in the process.

Kyle Michl: I love those examples. Shawna and I often think of the iceberg scenario around AI where everyone thinks about the model and the fun part at the top of the iceberg. But there's the full iceberg underneath the water of getting the data sets right and cleaning the data sets and getting ready for training. And I think this is a great example of the power of synthetic data, and I hadn't thought about the bias tangle. More curious if you have other thoughts and problems that synthetic data can solve.

Marc Bosch Ruiz: Yeah, certainly I agree on the accessibility part. That has definitely improved the ability of getting access to relevant data. I think the idea of control that that synthetic data technology brings is also very important, right? Like, you have control over how you want to design the data, how you want to generate that data so that benefits whether it's you're training your AI or evaluating certain things or just testing certain scenarios that you cannot necessarily have captured with real data, right? So that idea of control, but also it speaks to the biasing that the channel was talking about, right? Like the ability of removing bias or removing certain information from those data sets that they're not necessarily important to make certain decisions, but still preserving the value of that data, right? I'm thinking about medical records.
You can still try to build algorithms that can help diagnose certain diseases without necessarily needing to have private information about patients.

Kyle Michl: Got it. Thanks for that. So we've got these different use cases out there. We've got a better understanding of technology, some of the problems we can solve. So how are we seeing federal agencies using these synthetic generation technologies?

Shauna Revay: So when I think about the federal space, think about the scope of a lot of the problems that we're dealing with, there're things that really are large healthcare military that affect so many people, and so a lot of times these agencies do have access to some really rich datasets. But, of course, the problem is that a lot of it involves confidential or personal data, things that you wouldn't really want to distribute, especially not to the greater research community. And so that's another benefit that synthetic data provides is the ability to anonymize data so replicating real datasets with the same statistical properties with masking the actual people and the identities and the real datasets. So that's something here that the federal space can really leverage. I know one example during the beginning stages of the pandemic is that NIH had a large data enclave that was intended just for COVID research. And, of course, a lot of this had patient data, and so we needed to mask out those names. And so they actually created a synthetic data set and released it to the research community to help spur research with, you know, something that was a public health issue and really time-sensitive. So they were able to research things like risk factors, protective factors, long term health consequences, all of that. So that's like a really cool example, I think, of how the federal space can leverage this.

Kyle Michl: I love that and its examples. Anna definitely helps me visualize what you talked about earlier about the value of, you know, expanding those datasets. Thanks for that. Now Mark, I think you touched on and maybe show needed, as well a little bit of the social media aspects of this. Of course, there's a dark side to synthetic scenarios with concepts like deep fakes and other nefarious use cases. And our research found that just 35% of consumers are confident in their ability to recognize deep fakes. I saw it recently in America's Got Talent, which a little shame to admit that I watch it, but nonetheless it's not recently an example where

there were real time showing a deep fake of Simon Cowell singing on the screen and it was it was really compelling and you know, the fidelity was higher than I would have expected. So I'm curious, you know, how do you see deep fakes today? How convincing are they? And how quickly did it become an issue for public trust?

Marc Bosch Ruiz: I think they are certainly for the untrained eye they're starting to be at a point that they're extremely realistic and incredible. It's like unless you really know the flaws in advance of some of this technology it's starting to be very hard to identify that, right? I guess the good thing is that on our end is that you know like someone that has to create a deep fake, an attacker let's say, right? Like it has to get everything, right? Like the motion of the head of the mouth, of your expression, the shape of your ear, the symmetry of your face. Like there's a lot of laws that have to be preserved so that things look realistic. It only needs to happen that has one slot, right? Like it might just be one thing that doesn't look that natural to us to pick up that's a deep fake, right? So in that way we do have an advantage in identifying some of that but they're starting to become extremely hard to pick up on, right? Like I think we're trying to be at a point that, you know, like our brains really takes longer to make that determination whether something is realistic or not then we're willing to spend time on, right? And at that point it's a signal or an indicate mission that they're starting to be real, right? Like the 1210 millisecond that you're willing to spend trying to figure out if that's a real image or not. That may not be enough to really solve that that problem, right? And that's when we realize we need mechanisms that they're capable to do that. For us, right, to help us in analyzing the authenticity, the realism, veracity of something, right, because we are definitely not willing to put that energy in it just because of how real they look and feel.

Shauna Revay: And I really like your example, Kyle, because when you think about the America's Got Talent example, it was presented to the audience as part of the skit. So everyone knew to look, you know, to see if it was authentic and that was kind of presented that way. So, you know, maybe the video if you were studying it, trying to see if it was fabricated, you'd be able to pick something out. But when that scenario instead becomes you just scrolling through social media, are you going to be able to, you know, pick out specific videos when you're not looking for them necessarily, and I think we've seen this play out a little bit in real time with the war in Ukraine. I know at the beginning of the war a video was released of the Ukrainian president surrendering. Of course, it's fabricated so you can see the public trust issues that come up with an example like that, but things that, you know, are less scrutinized, they do, I think, have the ability to pull a lot of people just because they might not be analyzed to the level that they would need to be to be picked up as not authentic.

Kyle Michl: So it sounds like the public trust issue is real. And Marc, to your point, the amount of effort it takes to really determine what's happening out there is probably more than we're willing to spend. So, given these scenarios, what steps can we take to limit the malicious use of synthetic data, and what can we do to increase the public trust?

Shauna Revay: Yeah, I think, over time there's going to be this shift towards positive affirmation of authenticity versus taking it for granted. I think up to this point, it's kind of been the default, but I think you know, one step towards that is verifying authenticity. So there's tools out there like distributed Ledger technology, I think most commonly people. An example of that people would be familiar with would be blockchain technology are just ways to establish .

that provenance and authenticity of data on things that we're looking at.

Marc Bosch Ruiz: Yeah, I think, I think there's a component that will in my opinion help to gain public trust which involves showing that this type of technology has brings like very positive value to our and impact our society, right. Like you can pick many different technologies. The Internet, for example, right? Like I think we know of some of the upsides and the downsides that that technology can bring. But it's accepted for us because the positives really outweigh the negative, right? We can try to build technology and tools that can help us and be some sort of a, uh, first barrier of the sense. But ultimately, I think that if we can prove that this type of technology benefits society in in in ways that nothing else really can't, I think that's another. Direction where it can help you know with the building public trust and in organization, right?

Kyle Michl: Well, thanks, Marc. That makes sense. And we certainly started to touch on some of the evolution of the Internet when we talked about Trend One as part of the overall tech vision here. So I was thinking maybe we could pivot just a little bit. I think we'd be missing an opportunity to maybe take advantage of some of your experiences around AI and talk a bit about how that ties in here. So I'm just curious what other trends you're watching in the AI machine learning space that have the potential to disrupt government.

Marc Bosch Ruiz: I might not sound original, but I think like any technology that has the potential to. To reduce the dependency on real data, on data that we naturally collect or it's accessible to us. It's on my radar. I'm talking about AI algorithms that can, in a way, learn from the world the same way in the humans learn that not necessarily to show thousands of examples of one concept and then the moment you show 1000 examples of another concept, the previous concept is forgotten, right, so that ability of having machines that meaning more we learn as you humans, I think that's something that that I'm hopeful that that sooner rather than later we'll start seeing some of that and I think that's also going to have a profound impact on how we how we think about AI and sort of the end use cases that AI can you know facilitate for our societies.

Kyle Michl: I like that, you know, although I think some humans learn faster than others. Took me a while to figure out whether it was a muffin or a dog. Some of those pictures, some of your thoughts.

Shauna Revay: Yeah, when I'm right now. You know things that are on my radar when I look at the pace of innovation with some other generative models that have been released. I think that has the potential to be really impactful, especially since things are finally actually getting released into the open source. Also, I know we talked about this elsewhere in the tech vision, but the Metaverse and augmented and virtual reality, that really has a potential to change the way that we interact with the world. So I think there's potential for disruption there as well.

Kyle Michl: Got it. So I would hesitate to use the term exponentially, but they're really, you know? Tough to debate the fact that AI practices are really evolving rapidly. We talk about, you know, the trend year on year and the impacts that it's having given the challenges that federal agencies face and frankly a lot of the market is facing with regards to access to talent, technology gaps. For data use limitations, how can the federal government keep pace with this innovation?

Shauna Revay: I think it's important for agencies to just continue their commitment to research. I think this is the way that we're going to really be able to identify emerging trends, be able to apply it to specific federal use cases and really identify even risks with those technologies before they become systemic problems. So the research is really important. There and then also leveraging developments in that are released out on the open source and the research community that can really serve as jumping off points for innovation for these agencies to leverage.

Marc Bosch Ruiz: Yes I would agree with that I think it's very important to be at the sort of the source of innovation where research where the discovery with where, you know, like the proof of feasibility exists because many times it's in that stage where you're starting to envision, like the ups and downs of our technology, like the challenges, the potential use cases, right? And so the earlier you are thinking about some of the potential problems down in the future or potential application that that sort of this technology might may bring, this research may bring, I think it'll position decision makers in a match, you know, much better than something that it just took us all by surprise and nobody was really thinking about it, right? There's the program in the federal space that they're actively doing research on trying to understand, you know, like not just did 6 right but the core technologies that you can develop that yes you can use them to generate deep fakes but you can use them for other purposes and while you're working on developing this technology, you realize you're starting to think about the problems and you're starting to think about how can we sort of counter if it's sort of that technology becomes mainstream in a way that we don't necessarily expect, right? So that's one aspect. I think the other aspect for me is in terms of the Speaking of Australian gap, right, like we tend to think of AI as something that you only see in in in grad school if you're a computer scientist. But I think that, you know, AI has to be very—it has a lot of basic ideas that you know, like you can introduce. A lot of the curriculum way of I like we talk about middle school and certainly have our society be more familiar from the beginning on some of the how AI works we tend to think of a black box that we don't necessarily know what happens. It takes a bunch of data and it outputs something else. Sometimes we like that output, sometimes we don't like that output, but the black box in it doesn't need to be something that we wait until grad school to understand, right? Like I think that those are things that they think about in in a little bit 10 to 20 years down the road having definitely a workforce that that is very familiar and it's completely absorbed the ideas and the concepts of AI. I think that's also going to be definitely very helpful for us as a as a society.

Kyle Michl: Thanks for that, Marc, and I certainly agree. When, you know, I've spent time in some schools and different stem topics, it's amazing to me how ready you know the next generation is to absorb the newest technologies and agree very much that they're just natural absorbers of it and there's no reason we can't get them started early. But what I found interesting is you both talked about the importance of research as a way to kind of fuel innovation. You know, we've spent a lot of time with clients also on the potential challenge of moving from that research phase into, I'll call it the pilot or pragmatic implementation phase. I was wondering if you guys had a few thoughts on some of the challenges or some of the benefits of moving from that research phase to that implementation and whether or not that's an important part of keeping the innovation lab, the learnings, the proof points associated with getting into the use cases or testing those hypotheses in the real world?

Shauna Revay: Yeah, absolutely. And I think there's been a big push of standardization of kind of these processes over the past couple years to really help us take those proven concepts that we might be developing in research and build those into productionized models. So there's open-source libraries, products you can. There's a lot of tools out there to help make that leap between research and implementation. And I think federal agencies can definitely benefit there by making their own process that works from them, using their tools that they're familiar with to help bridge that gap.

Kyle Michl: So Marc, you two guys both talked about research, and what's interesting to me is, you know, research is in a core part of enabling innovation, one of the other things that we've often seen with our clients. Is the power and value that they get when they can get to a successful pilot or prototype where they can test some of those research hypotheses in that real world? Is that something that you're seeing as being important for maintaining that innovation kind of pragmatism?

Marc Bosch Ruiz: Yeah, absolutely. I'm seeing some research programs because we often think of or we found ourselves in all these research programs that they're like a valley of death where a lot of good ideas they just don't make it past the proof of concept in a lab setting to the applicability, right? And I think there's a lot of opportunity on setting up programs that the whole technology development. So the entire patent, their pipeline, sort of gets carried through like I'm talking about like the basic research but then starting to like be able to like see how far we can take these type of prototypes and start involving end users and others that ultimately will become the users of that technology closer together right? Like from the requirements from data available to test in relevant settings some of these techniques but also at the same time continue to like fuel that basic research where a lot of the ideas come, right? So, being able to have research programs that can sort of cover that full spectrum, I think it's going to be necessary and very important with moving forward particularly with technologies like. Like, yeah.

Kyle Michl: Marc and Shauna, thank you for joining us today and sharing your insights. You were fantastic. Yeah, thank you. I'm fine. And thank you for joining us today. A reminder that our next episode of the federal catalyst will focus on a trend for computing the impossible. Join Accenture Federal Services CTO Chris Copeland for a thought-provoking conversation on that trend. Finally, I encourage you to explore the Accenture Federal Technology Vision, if you haven't already. You can find it at www.accenturefederal.com. I am Kyle Michael, and it was my pleasure to be your host today. Connect with me on LinkedIn and let me know what you thought of today's show.

Announcer: This was the Federal Catalyst with Accenture Federal Services. If you liked what you heard, subscribe to your favorite podcast provider, and share with your social networks. Questions? Contact us at hello@accenturefederal.com. You can find us on the web at accenturefederal.com and accenture.com/federal. Until next time, thank you for listening to the Federal Catalyst with Accenture Federal Services.