MIT CSAIL Alliances | Arvind_Narayanan_Project_4

Welcome to MIT's *Computer Science and Artificial Intelligence Labs Alliances* podcast. I'm Kara Miller.

[UPBEAT MUSIC]

On today's show, AI is a game changer but that doesn't mean it can solve every problem.

Really, it doesn't seem like there is much signal in the data beyond saying someone who's been arrested before is more likely to be arrested again. In my view, that is the entire logic of this tool and it's dressed up in this aura of machine learning and AI.

Computer science professor and popular newsletter writer Arvind Narayanan joins us to explain how to separate hype from reality and how to truly understand how the landscape is shifting.

You don't have to be, quote-unquote, "on top of things." You just have to pick up foundational knowledge so that you can have a general understanding of the direction in which things are going and where things are. And that is not going to come from the news.

That's all coming right up. In business, in government, in nonprofits, the opportunities to use various types of AI are multiplying. One of those is predictive AI, which has been used in the US Justice System to essentially try to figure out what's going to happen next.

It's used in predictive policing, that's a big area. And then when someone is arrested, should they be detained until their trial, which could be months or years, right? Or should they be released on bail. And then for parole, sometimes even for sentencing, these are all places where the system needs to make a decision. And there are many reasons why AI might seem very appealing here.

Narayanan is a professor of computer science at Princeton University. He is the director there of the Center for Information Technology Policy. And as he noted, the rationales for turning to AI, they really go on and on.

We have this enormous problem of prison overpopulation and jails. It's a human tragedy. But also from the systems perspective, they would like to decrease the jail population if they can do so without a cost to public safety. There is the issue of people being a flight risk, not appearing for a trial. And in addition to efficiency, there is the issue of bias in the criminal justice system. There's a lot of research on how judges can be biased in their decision-making.

Not surprisingly, then, corporations have come up with products to help those in the criminal justice system predict the future. One piece of software that's been commonly used is called Compass.

What it aims to do is produce risk scores in various categories. And these risk scores can be used at many points in the criminal justice system. But let's assume someone has just been arrested and the judge needs to decide what to do with that person.

So will they be detained without bail, or will bail be set, or will they be free without bail pending their trial. So those are some of the possible decisions. Another decision is to have them go free but have ankle monitors or another kind of surveillance. And what Compass tries to do is provide an algorithmic score that can help the judge make this decision.

He says the software essentially gives judges a bunch of probabilities. What is the probability that someone will commit a crime in the next couple of years? How likely are they to not show up for their trial?

And the way it makes these decisions is it is trained based on a data set of past offenses and past defendants. And the information it has about a defendant includes various things about them like their age and the number of prior offenses. But also, there's a questionnaire consisting of 137 questions, apparently. And it includes some weird ones, things like how often are you bored. And [CHUCKLES] supposedly the claim is that this algorithm is so good that it can use these seemingly irrelevant factors to predict these risk scores.

But Narayanan says the AI often falls short. A few years ago, ProPublica, a nonprofit that does investigative journalism, argued the tool was racially biased. Narayanan notes it tended to label black defendants high risk more often than white defendants. And when the judges released those riskier defendants, it turned out the false positive rate was very high. That's the number of people who are labeled as likely to reoffend but don't actually go on to reoffend. Even beyond bias, though, there was a deeper question.

Do these predictions even work? When we looked at those numbers, we were shocked. The way the accuracy is measured, it's by a metric called AUC. The technical details don't matter but the bottom line is, the baseline is 50%. Random guessing would give you 50% accuracy. And Compass and other tools generally have an accuracy of less than 70%. So this is slightly more than the flip of a coin. And there are many reasons why even that accuracy might be exaggerated, might be an artifact of the way that it's measured.

So really, it doesn't seem like there is much signal in the data beyond saying someone who's been arrested before is more likely to be arrested again. In my view, that is the entire logic of this tool and it's dressed up in this aura of machine learning and AI.

And in an age when AI is indeed fundamentally changing companies, and how we work and manage data, Narayanan says this criminal justice software is definitely not alone in using buzzwords to try to sell itself. It's a topic that Narayanan dives into as the co-author, along with Sayash Kapoor of a brand new book, *AI Snake Oil, What Artificial Intelligence Can Do, What It Can't,* and *How to Tell the Difference.* And by the way, he does believe that there is a role AI can play in the criminal justice system.

It can be used to analyze the behavior of judges, find particular ways in which judges are biased. And then we can use moral reasoning to think about whether that's actually bias or not and give that feedback to the judges to help them do better. That's one way it can be used.

Another way it can be used is just basic statistics. So if the only signal that this tool is extracting is that people who have been arrested before are more likely to be arrested in the future, just give that information to judges. Let them just, look at it by hand so that allows them to preserve their autonomy instead of turning the decision over to this black box system.

So does that mean when it comes to predictive AI, in particular, that we should always be cautious? Narayanan says, well, yeah. It just doesn't work that well.

It's hard to predict the future. Whether or not we use data, whether or not we use AI. My colleague here, Professor Matt Salganik, who's a sociologist, and I have been coteaching a course called limits to prediction. And so we looked at something like a dozen different domains to see how well AI can predict the future, whether that's weather or people's behavior.

And it turns out that across the board, as far as we can tell, there are strong limits to predictability. It's hypothetically possible that if we collect much bigger data sets in the future, we're going to be able to do better. But based on what I'm seeing so far, I would guess that we're going to see maybe small increases in accuracy. It's not going to be a fundamentally different picture.

So why is that? Matt, in fact, led a project called the Fragile Families Challenge, which was a big machine learning collaborative effort with 160 different teams who were trying to use a data set of kids' life outcomes. These kids had been followed from birth up until age 15 or so, several thousand kids. And this kind of rich, detailed data set with 13,000 variables per child or per family was somewhat unprecedented in social science at the time.

And it turned out that the reason why machine learning wasn't able to do much better than random guessing or simply guessing based on people's socioeconomic status and other background variables wasn't because of limitations of how clever these algorithms are. Because very different kinds of algorithms were all converging on the same sorts of predictions, which were only a tiny bit better than random.

It just seems like [CHUCKLES] there should be common sense. But based on the data as well and based on interviewing a lot of these children, there's just so much going on in these children's lives. And one interview, it turned out that a kid was really struggling and there was a kind neighbor who talked to them every day, fed them blueberries. I forget the exact details. But the presence of that neighbor really helps this child do much better at school and in life, et cetera. And we're never going to have a data set that codes for the presence of a blueberry-feeding neighbor. That's just a fundamentally random thing. I don't think it's the kind of thing we're ever likely to be able to pick up from data sets.

Yeah, I mean, exactly. I remember that example from the book. And it just goes to that point that you made of like, it's very hard to predict the future. Somebody moves in next door. The person could be an incredibly bad influence or a retired biologist who tutors you in your biology homework in which you weren't doing very well. And then all of a sudden, you are. And I mean, who can predict that, right?

Exactly.

So if you're a company-- I feel like so many companies look to predictive AI to solve a very reasonable and hard problem. I mean, I think about hiring. If you're hiring at a very large company, you're fielding thousands or tens of thousands of applications. It might feel totally not feasible to assign humans to deal with all this stuff. So you're looking to something smarter.

So do we have to just say forget predictive AI, forget machine learning here, we can't use anything to weed through-- as we were saying with the court system, when you have so much incoming traffic, the people need some help to figure out, well, what should I focus on. I can't focus on everything at the same time.

Right. For sure. So the message is not to give up. But I'll say a couple of things. One thing we've noticed over and over again and hiring is a great example is, companies are desperate to turn to AI when there is something broken basically in their organization about certain processes. And yeah, hiring, again, is a perfect example because companies are getting hundreds, maybe thousands in some cases of applications for a single open position. And it's just not possible for people to wade through all that.

But I think we have to start with the recognition that there is something broken here and we're not necessarily going to fully fix it with technology. In fact, what we're finding is that as companies turn to predictive AI, candidates are turning to AI to increase the number of jobs that they can apply to by having automated ways to submit their resumes.

So it's just a spiral, right? The problem is getting worse. And more technology is just leading to an arms race on both ends. And so there has to be a different approach, I think, to weeding out those applications. And AI is not going to be the answer there.

That said, I think AI, and data, and algorithms can be very helpful here in many ways. Instead of delegating the decision-making to AI, it can be used to analyze the past performance of candidates and identify the kinds of qualities that turned out to be a good fit for the role and so forth. And that can help inform the hiring process.

So AI for analytics is very different from AI for decision-making. And the former, to me, seems much more justified than the latter. Because there's going to be a human in the loop. Whereas when you have AI involved in decision-making, the fine print always says, oh, AI shouldn't make the final decision. There should be a person responsible.

But it turns out when we look at what actually happens, that person is just there to rubber stamp the AI decision and it takes accountability out of that process and removes the ability of the candidate to present themselves, put their best foot forward, have their case be heard by and evaluated by a person. I think those are important aspects of the process for preserving dignity and other things that are really hard to quantify using data and machine learning. So yeah, again, I think there's an important role for AI and machine learning but I don't think it's automation.

So just to stick with AI for hiring for a second because hiring, certainly, is something that every company thinks about, struggles with. When you talk about not having AI make decisions but maybe thinking about what has worked, would that mean saying, OK, well, these are the kinds of people who've really done well in this company and having some sort of AI analysis essentially done on their background, like, oh, it turns out that people who've had classes in logistics, they tend to do the best because they really understand it, is that the kind of thing we're talking about?

That's right. And again, we have to be careful here not to just throw AI at the problem and just trust whatever comes out because it could just be picking up biases in the data. So one example would be that AI might predict that people who are good at chess did well in software engineering positions. But that might not be because chess is causally making someone better at software engineering. It could be because in colleges, people who are good at chess are more likely to be socialized into taking computer science classes, for instance. So it could be something like that which is not the kind of thing you want to act on.

On the other hand, it might tell you something really important about your company. It might turn out that the fits between the particular manager and their managing style and the employee is the most important determinant of success. And so data, I think, is incredibly valuable for those kinds of insights which can be actionable but are different from automated decision-making.

You write that you personally do use AI a bunch in your work. Presumably, you use it in ways you think is helpful as opposed to harmful. How do you use AI in a beneficial way?

Sure. So let's talk about generative AI because that's the more consumer-facing part of AI. And we've been talking about predictive AI so far where I think there are fundamental limitations. Generative AI, on the other hand, what we say in the book, is that this is a genuinely new technology, whereas predictive AI for the most part is just century-old statistics dressed up in new clothes.

Generative AI has had a lot of innovation behind it. It is capable of genuinely new things in many areas. There are certainly dangers. There's a lot of hype. But on the other hand, I think most people and most companies can find productive uses for it.

So in my own work as a researcher, I think chatbots and generative AI are helpful in many, many ways. Let me just give a couple of examples. One is for automating certain mundane tasks. So I had a bunch of links that I had collected for an article that I was writing, for instance. Turning those links into properly formatted citations, especially in LaTeX, which is a powerful but kind of annoying software that a lot of us use in academia. Super annoying. Very time consuming.

There is existing software for that. But it turns out that a 10-minute script that I wrote using GPT-4 was able to do much better than this existing software that has been developed with hundreds of hours of software engineering effort. So that's really cool, to be able to automate a mundane task with a very minimal amount of effort. I found one particular use for that that's not going to be relevant to other people unless they're also writing research papers for a living. But I think everyone can find something like that probably. And that's on the mundane end.

On the more exciting, for instance, when I'm trying to learn a new research topic, I find that chatbots can be one very useful tool, one among many. I haven't stopped using books to learn new topics. But I can't ask a book a question. I can't tell my book, here's how I understand this. Does this make sense or are there gaps in my understanding?

And a lot of the time when I want to learn a new thing, I don't even know what are the right technical terms to Google to find relevant papers on that topic. But I can describe a vague sense of what I'm looking for to ChatGPT and it will tell me what the terms are and can even potentially automate the search for finding papers. So there are pitfalls. But at the same time, this can be a really powerful tool for knowledge workers.

It also helps you locate things quickly. I mean, I've sat in front of books before knowing that I wanted to find some particular thing but I had no idea how to find-- I mean, it wasn't in the index. I had no idea how to find it in the book. But if you know enough of the pieces, you could just ask ChatGPT, or Gemini, or whatever, can you tell me whatever it is. What happened in 1955 and not have to search through 300 pages of a book to find it.

Exactly. And let me tell you a funny story about that. So I have a draft of our whole book sitting inside ChatGPT. And I find that very useful. Often, someone will ask me a question, did you write about this in the book? And there's 300-plus pages in the book. We've written so many other things as well. Did I write about this particular incident where AI went wrong? I'm not sure.

So very often, the way I answer that question is by asking ChatGPT and it uses something called retrieval augmented generation, which is a smart kind of search, which Google has also been using. Again, there are pitfalls. Google has been in the news for bungling a lot of that and giving answers from *The Onion* as if it's a correct response to people's questions. But again, keeping all that in mind, it can nonetheless be a useful tool as long as you're willing to put in that little bit of extra effort to verify the answer that AI is giving you, knowing that it can make stuff up.

Right. I'm sure people talk to you all the time about implementing AI at their companies. I wonder if you can talk about a way or two that you've seen AI most fruitfully implemented at companies. We've talked about some ways that maybe it doesn't work so well. But are there fruitful implementations that you can point out?

Yeah, definitely. I maybe won't give specific examples but I'll talk about general trends that I've seen. So when we look at generative AI, I think, as we've been talking about, I think most knowledge workers can make use of this. I think it's really helpful for companies to have an actual strategy instead of-- or in addition to just letting employees explore on their own. I think that's also important.

But they need to feel empowered. They need to feel like that exploration they're doing in order to figure out how they can use this technology is supported by the company, not something they're doing on their own time, ways internally for people to share expertise with each other on what works, what doesn't. Because that's going to vary a lot from company to company.

So there are efforts like that at my university, for instance. Because the specific kinds of pitfalls that professors face when using generative AI, whether it's for research or for teaching in some cases, that's very different from what a typical company will face. And what one company will face is different from what another company will face.

So having those efforts that combine a bottom-up exploration with a top-down way to have information sharing, quality control, privacy and security policies in place so that data doesn't leak to AI companies, especially sensitive proprietary data. Teaching people about the pitfalls, what to avoid, and helping people feel empowered. So I think those are all things that are super helpful.

When it comes to predictive AI, I think most companies can benefit from some kind of analytics. The one thing I'll say here is that something I've found over and over is that these off-the-shelf solutions, whether it's for automated decision-making, which, again, I think there's a lot of hype there, but even for analytics come with a lot of drawbacks. Come with handing over a lot of control to the AI vendor. And at the same time, having a tool that is not really tailored to the needs of the company.

I think people have an instinct to try and go buy AI from an AI vendor because of this myth that it is this super sophisticated thing that only a few geniuses have mastered. Whereas, in fact, a lot of analytics, a lot of analyzing the data that a company might have is statistics and basic machine learning models that can easily be built in-house. And when it's built in-house, it can be built in a way that's much more tailored to the needs of the company and in a way that is much more attuned to the potential pitfalls and the potential downsides that can come about.

When you think 5 or 10 years down the road, do you imagine that AI is going to reshape the way most companies do business? And if so, how-- if you do think that, how do you imagine that looking?

I think AI is going to have an impact on almost every business. What that impact is, how big it is definitely depends on the sector and the specifics of the company. There are certainly some sectors that are already seeing a huge amount of impact. So we're seeing, for instance, translation, which is one of the areas where generative AI makes automation relatively readily possible. Obviously, we need people reviewing AI outputs. But AI can do a pretty decent first pass with translation. So industries where that automatable work is a big part of it are going to see big impacts.

In other companies, there are going to be smaller but I think important impacts. So one thing that I've learned from economists is that AI automates tasks, not jobs. And jobs are bundles of tasks. And so I think in many different jobs, certain tasks are going to get automated or partially automated.

And I think what's going to be critical is whether a company is able to allow its workers to upskill so they take advantage of automation in useful ways, as opposed to being surprised when AI is able to do a new thing that it wasn't able to do before.

So being proactive and planning for the possibility, the probability that every couple of years we're going to have important new AI capabilities that companies will need to react to, as opposed to seeing this is a one-time thing. Oh, we didn't have ChatGPT and now we have ChatGPT, let's adapt to this, as opposed to how do we reorient ourselves so that we can keep adapting as AI capabilities keep improving.

Right. That's really interesting. You also have a great-- you were talking before about this idea of AI not automating people but automating tasks. And you talk in the book about when ATMs came in. And I feel like when I was a kid, there were a lot of bank tellers. I remember my mom standing in line at the bank teller. But of course, almost nobody does that anymore. People go to ATMs. But then the punchline is, what happened to bank tellers?

Right. So the really surprising thing is that automation actually increased the number of people who are employed as bank tellers. And the funny reason for that was that automation made it cheaper for banks to open branches. They opened many more branches. Now you need fewer tellers per branch, but not zero, because not everything that a teller does is automated. There are still many things that require human interaction. So overall, when you factor in the increased productivity of bank tellers, but at the same time, the fact that there are many more bank branches, it turned out that the demand for tellers actually increased.

Give me your take on the big company versus small company struggle here. I've heard people say that small companies have an advantage in some ways they've never had before because they've got AI on their side. And that gives them this multiplying power.

But then you've also got people who make the very reasonable argument that, look, all this compute power is very costly and who do you think can afford to pay those costs? It's very big companies. So in that small company versus big company fight, how does AI change the game?

And I can only guess here. Who knows what's going to happen. And certainly it's going to differ industry by industry. But I actually think in many cases, it's both. It's going to make certain big companies even bigger. The tech giants, notably, who are the ones in control of the leading foundation models and can charge rents, so to speak, in econ speak, when the rest of the world needs to use those models in hundreds or thousands of different products.

But at the same time, it's also going to allow small shops by taking advantage of automation to do things that they couldn't do before at the scale of five people or whatever it is. And I think we saw the same thing with the internet. It enabled a lot of small businesses. But at the same time, it created some very big companies at the top end. So I think both those groups are right, ultimately.

It's such a good point because if you make candles and you want to sell them on a website, your ability to create a website without knowing code or anything is so incredible. But as you say, clearly, we have some incredibly powerful companies out there. So the empowerment of the small business didn't exactly detract from them.

That's right.

I wonder why it is you think that people tend to buy into the hype around AI, what you see as the hype around AI? And how somebody who really, really wants to understand AI, integrate it into their company, how they should follow the news and follow progress without getting sucked in to things that don't, at least in your opinion, make any sense?

[CHUCKLES] Sure. I mean, AI hype comes from so many sources Most obviously companies. It is very profitable for them to sell AI as something more than it is. To just rebrand whatever it is they've been doing as AI and actually, in some cases, claim that their product is AI when it's just hidden humans behind the scenes. So there's a lot of that going on from companies.

I think for the media, I think there are a lot of incentives to hype things up. It generates clicks. And I think there's this troubling aspect of access journalism where companies-- pardon me, where journalists by not being too critical about AI hype and just generally going along with how companies are portraying their products can make sure that they have early access to things, are able to access people inside companies to get quotes and so forth. So there's this mutually reinforcing cycle.

But let's also not forget researchers in academia who also have so many incentives to hype up things. And a lot of our research and some of the parts of the book are about how a lot of AI hype is coming from researchers. And I think, again, all of these reinforce each other. And it can be hard for everyday individuals to tell what's real and what's not.

So how to stay sane in all of this? I think following the news might not necessarily be the best approach. [CHUCKLES] I think things actually don't change as fast as they're often portrayed. So one thing I've been pointing out recently is after ChatGPT came out, I mean, that was, for a lot of people, their first introduction to chatbots. And then there was this assumption that every three months, there's going to be a bigger model that's going to be able to do a whole new set of things. But that just hasn't happened.

So in many ways, you know, if once a year you devote a week to learning what happened that year in AI, you're not going to be too far behind. I'm not saying you should only think about AI once a year but I'm just making the-- [CHUCKLES] making the point that it's-- you don't have to be, quote-unquote, "on top of things." You just have to pick up foundational knowledge so that you can have a general understanding of the direction in which things are going and where things are. And that is not going to come from the news.

So that's going to come from things like lectures, or books, or tutorials, or things like that. And even when it comes to generative AI, playing around with the technology by yourself. I think building up that foundational knowledge is going to be much more valuable than trying to stay on top of the news.

Arvind Narayanan is a professor of computer science at Princeton University. He's the director of the Center for Information Technology Policy and the co-author of the forthcoming book, *AI Snake Oil, What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference.* Arvind, this was great. Thank you so much. I really appreciate it.

Thank you for having me on.

[UPBEAT MUSIC]

And a quick note before we go about an upcoming course from MIT CSAIL on cybersecurity. As companies become increasingly digital, cybersecurity has become a top concern for business leaders. So join CSAIL for the course, Cybersecurity for Technical Leaders. It's tailored to give executives a deeper understanding of contemporary cybersecurity issues. Businesses can upskill their leadership and keep enterprises secure.

Listeners to the podcast can save 10% with the code MITpodcast. The course has been developed by MIT CSAIL and MIT PRO and is delivered by Simplilearn. If you'd like to know more, just email us podcast@CSAIL.MIT.edu. That's podcast@CSAIL.MIT.edu. I'm Kara Miller. Our show is produced by Matt Purdy, with help from Audrey Woods. Tune in next time for a brand new edition of the *CSAIL Alliances Podcast* and stay ahead of the curve.