Detecting the Good and the Bad With AI: Airbnb’s Naba Banerjee

SHERVIN KHODABANDEH: Machine learning can detect fraudulent activity, but it may also miss good behavior. Find out how one platform company manages these risks and trade-offs on today’s episode.

NABA BANERJEE: I’m Naba Banerjee from Airbnb, and you are listening to Me, Myself, and AI.

SAM RANSBOTHAM: Welcome to Me, Myself, and AI, a podcast on artificial intelligence in business. Each episode, we introduce you to someone innovating with AI. I’m Sam Ransbotham, professor of analytics at Boston College. I’m also the AI and business strategy guest editor at MIT Sloan Management Review.

SHERVIN KHODABANDEH: And I’m Shervin Khodabande, senior partner with BCG and one of the leaders of our AI business. Together, MIT SMRand BCG have been researching and publishing on AI since 2017, interviewing hundreds of practitioners and surveying thousands of companies on what it takes to build and to deploy and scale AI capabilities and really transform the way organizations operate.

SAM RANSBOTHAM: Today, Shervin and I are speaking with Naba Banerjee, director of trust product and operations at Airbnb. Naba, but thanks for joining us.

NABA BANERJEE: Thank you; a pleasure to be here.

SAM RANSBOTHAM: I think most people know Airbnb, but maybe tell us a brief overview of the company and what you do.

NABA BANERJEE: Airbnb stands for a place which fosters connection and belonging. Millions and millions of hosts open up their homes to complete strangers, and you and I and our families get to travel to an exotic destination, and instead of staying at a hotel, you actually get to truly immerse yourself in the local culture. And at times, if you’re really lucky, you get to stay at the house of a host and actually experience hosting as it’s meant to be, thereby creating and fostering connection and belonging.

And in this world that is increasingly becoming insular and connection is becoming a rare thing, I am so glad to be part of a company that is trying to create more connection. And what my team does, the trust and safety team at Airbnb, is along the entire customer journey, right from the time that you create an account on Airbnb throughout the journey. At times, it is possible that something could go wrong: A fake account could get created, or your account could get taken over. The listing you’re trying to stay at could be a fake listing; the reviews you’re looking at may not be exactly genuine. But our job is to make sure that you can focus on your magical stay, the host can focus on being the perfect host, and we try to anticipate some of these risks, and we minimize the
risk of those kinds of issues from happening. And using technology and data to enable that magical user journey is what my team does.

SAM RANSBOTHAM: I think we all hear headlines where something goes wrong. How much goes wrong in the course of the gazillions of transactions that you do?

NABA BANERJEE: That’s a great question. In 2021, we had around 70 million trips that happened on Airbnb, and less than 0.1% of those trips resulted in a host or a guest reporting potentially an issue that happened in that stay. And when our team went in and investigated and actually looked at where the real harm happened or somebody had to be removed from the platform, that number is even smaller.

I know that even one bad incident is probably too bad and should not be happening, but in the bigger scale of things, what you come back to realize is the majority of people are actually good. Ninety-nine point nine percent of people are meaning to just go about their daily business, enjoy a great stay, or be a great host. But sometimes, [on] that very rare occasion when that bad thing happens, we want to make sure that we reduce the risk of that.

If you really look at where a person is causing harm to another person, those incidents are like a fraction of a fraction of a fraction. But when we talk about safety, we’re looking at [things like] a carbon monoxide gas leak. It could be slip and fall; it could be property damage, where glass broke and you probably had a stain on your carpet. We are also looking at those kinds of issues.

SHERVIN KHODABANDEH: What I’m hearing from what you’re saying is, your role is, No. 1, a good role. You’re the protector of the good and the “warder-offer” of evil, right?

NABA BANERJEE: I love that. You should be part of our branding team.

SHERVIN KHODABANDEH: It’s a good thing, right? It’s a very purposeful role. And the second part of it is, you’re talking about microscopically small incidents.

NABA BANERJEE: Yeah, needles in a haystack.

SHERVIN KHODABANDEH: Right? And those are fun problems. Tell us more about under the hood — what happens?

NABA BANERJEE: I would say it’s three parts. One is, we start with the “why,” which is — you framed it really well — that we are the protector of the good and trying to ward off the bad. We can never say that we stop everything bad. Like, to Sam’s point, you cannot do that. But how are we learning and getting better? And that’s where AI and ML [machine learning] comes in, where [we ask], “How can we use [the] power of data and technology to learn and get better?” And every year, the risks don’t stay the same; the data keeps growing.

And then the second thing is, “How do we use technology responsibly?” Which is, you could go too blunt and try to block as much as possible to specifically look at fraud and
safety incidents and keep them down, but then, if the data that you are using to make these decisions has bias in it, is potentially creating unfairness, then you could close off your doors to a lot of good people who genuinely are just trying to have a good stay. So [it’s about] using technology responsibly. ... It is having self-governing mechanisms, which is, we have our privacy team, our data privacy team, our infosec team, our anti-discrimination teams, which are ... almost like mirrors to the trust team. [There are also] other teams in Airbnb, where every time we are really laser-focused on this particular fraud vector, this bot detection, this account-takeover detection or party detection, they are constantly making sure that are we using the data in a safe way and a transparent way, giving users control and then not introducing bias.

And then, lastly, having ways for us to give people a path back, which is, we are going to — based on the data available to us, and the technology available to us, and the maturity of that technology — make our best possible decision. But if we get it wrong, can we create transparency and a process for users to appeal and get back on the platform when we do block them or remove them so that, again, we can learn and get better? So that, at the highest level, is how my team is working under the hood.

SHERVIN KHODABANDEH: And what are some of the use cases?

NABA BANERJEE: One of the use cases that we have talked about quite a bit is, you know ... I’ll take you on a little story, which is, when I joined the trust team in 2020, the world had just shut down; the pandemic was raging. And very similar to how the world was waiting for a vaccine, while waiting for a vaccine, we went into lockdown mode.

That was a blunt instrument ... to prevent people from exposing themselves to situations which would make them unsafe. Similarly, Airbnb became an unfortunate product-market fit for unauthorized house parties. When the hotels and bars shut down, unfortunately that less than 1% or 0.1% of users who were looking for places where they could throw these parties, they started renting out Airbnbs. And as a result, we realized that it was causing a lot of pain to our communities of hosts and guests who use this platform, which was based on a foundation of trust. So initially, we had to use some very blunt instruments.

We first instituted our global party ban, saying that Airbnb does not allow for any kind of party whatsoever. Second, we started looking at some patterns and found that someone under 25 booking an entire home for just one night or two nights less than 50 or so miles away from their home — those seem to be all signals that lead to parties. When parties happen, we usually find ... that account was created yesterday and there are no positive reviews at all. Then we implemented a blunt rule called “under 25,” and we put that in place. Now, we had to be careful, because those rules don’t exactly work globally. But that worked in North America, and we saw right away that there was a reduction in the number of unauthorized parties.
But over time, just as it happens with rules, we saw that people started to game those rules. They would get an older friend to book the reservation for them. And instead of one night, they would book it for three nights, and so on and so forth. And the impact on good users was so high that we said, “We kind of need to move to machine learning. We need to get smarter.” And so we did our first pilot in Australia, where we segmented the place and looked at the place that had this new party model that we’d built versus the other. Do we see a reduction? And again, this is like looking for a needle in the haystack, but we still saw 35% fewer parties in the area that had the party model versus not. And when we started experimenting in North America, we did see that it was as effective as the heuristic, with lesser impact to good users.

And that kind of gave us the confidence to keep moving forward, because not doing anything was not an option. We had to do something and had to be smart about it. That was one of the use cases where we could go out and talk about it — not claiming that we are stopping all parties, but at least Airbnb is taking a strong stance about it.

SHERVIN KHODABANDEH: One common threat that I’m hearing in how you’re describing the problem and the solution is a need for constant adaptability and ongoing learning, and using the right instrument for the right job. At times it’s a rule, then maybe its unsupervised learning, then there’s a human intervention. It does sound quite multidisciplinary, including the part where the rules change, or the fraudster’s mode of operation changes, and you also have the issue of “Is it too blunt? Does it limit some unintended but good or benign behaviors?” How does the human and machine collaboration or interaction work in these situations?

NABA BANERJEE: You summarized it really well, Shervin, which is, it’s not a one-and-done thing. You build it, and then you have to constantly learn and optimize from even the kind of decisions you are making, as well as how the world around you is evolving. And that is an area [that] humans are really good at, which is, you do have to have the discipline of training your models with the latest data, with the right data, making trade-off decisions on what is the optimal threshold at which you are going to say, “OK, this is risky, and we are ready to use that risk.”

At Airbnb, we have been constantly on this journey of how to leverage humans in the loop. I have a trust operations team in addition to the product and planning teams. This operations team constitutes full-time employees as well as agents globally. And whenever we build a new model, in the beginning we often have the model make a risk threshold-based decision: “This is the top 1%. This is the next tranche. This is the next tranche.” And at times, we will route that to human beings to then create labels on top of that to say, “I agree with the model’s decision,” or, “I don’t agree with the model’s decision,” which then serves as training data back to the model.

And then, when we have more confidence in the decision-making capability of the model, we would go toward auto-decisioning. This party model, to your point,
right now it’s completely auto-decisioned. But what ends up happening is, when we get an appeal back from a user saying, “I was incorrectly blocked,” that then serves as data that’s coming by the customer telling us. And then an agent looks at that appeal and says, “Actually yes, this decision was incorrect.” And if the agent made a correct decision, that customer appealed correctly, and no incident happened, we know that that was a false positive.

But let’s say the agent approved, and then that customer went on to throw a party; that was a false negative. We are learning from that continuously, but we do it iteratively. First, we try to have the model direct to a human. And again, there is no guarantee that the human is better at decision-making than the model, but we constantly measure the performance. Then we go toward higher-order decisioning [with less] human decisioning. And sometimes we will also have … let’s say the party model makes decisions. People go about their merry way, have the reservation. Sometimes after the reservation has been booked, we will also still have humans look at the top 1% risky reservations to try and correct anything that they can.

So it is either trying to balance that out as to, when are humans the right people to make the decision? When should the machine make the decision? It’s been a not-easy, nonlinear journey that we have been on.

SAM RANSBOTHAM: You mentioned at first you’re maybe reviewing more cases and then learning from those cases. But at the same time, you have people … these 70 million transactions are happening constantly. And, for example, when you gave the illustration of the hotel … shutdown at the beginning of the pandemic, that’s time where it’s happening to individual hosts and customers. How do you balance that trade-off of needing to get something in place quickly [while] at the same time needing to go through a process?

NABA BANERJEE: That probably is the hardest thing, and the hardest thing in my entire career I would say that I have dealt with. Before this, I was managing all of product for Samclub.com, which is part of Walmart, and I thought my job was hard then. And while we were serving customers, helping them check out, it still felt like we were on a growth path, and we could choose how fast we grew or how slowly we grew. And I always had teams that were looking at people who were trying to check out [but] couldn’t and serving them while we were building longer-term architecture, but people’s lives were not at stake; [it was] people’s money — fraudsters using stolen credit cards and swindling the company of millions of dollars.

There is a team at Airbnb that is trying to protect the good and ward off the bad. What we do is that you have to get really good at prioritization, which is also very hard to do in the world of trust and safety.

SHERVIN KHODABANDEH: And in many ways, I mean, the choices are what, Sam, you said in the beginning: These kinds of economies are not risk-free. They cannot be risk-free. And the good collectively outweighs the evil, so then it becomes all
about prioritization. But I have to say, we’ve talked to many people in analogous roles that you have, in other organizations — you mentioned Sam’s Club, others — where they’re using technology, digital AI/ML to drive a variety of use cases. I think yours stands out in the sense that it is a continuous balancing act, and the stakes are high. And then you have this sort of thread of purpose and the humility of saying, “Well, we believe that we’ll make mistakes, and the system cannot be mistake-free.”

NABA BANERJEE: Yeah.

SHERVIN KHODABANDEH: But over time, things are getting better. Which brings me to my next question: How do you measure this, your effectiveness? Is it in terms of number of incidents, and some severity and frequency of bad things happening?

NABA BANERJEE: I’m glad you asked this question, because I was naturally going to go toward this, which is, at the highest level, the metrics that we measure ourselves on are number of fraud incidents per million trips, number of safety incidents per million trips, as well as good user impact. That’s the balancing metric to see “How many good listings did we block? How many good hosts and good guests did we prevent from moving forward to truly understand our false positives and false negatives?” But at the end of the day, it’s good trips that we are looking at across the board.

We are also looking at dollars in terms of fraud loss that is potentially happening. We are looking at customer support tickets that are coming in, user NPS [Net Promoter Scores] resulting from anyone who encounters a friction that is thrown by our teams. So at a basic level, we are looking at these metrics. But one of the challenges that, for the first time, again, in my career I’ve run into — not so much in the fraud world, where we can A/B test — we are constantly iterating our models.

SHERVIN KHODABANDEH: You are probably also, when you do that, preventing quite benign behaviors and intentions as well. But as you said, there is a point where the human judgment needs to trump and say that this is something where even the risk to one person might be too high. And yes, I mean, we always talk about exploration and exploitation when it comes to machine learning, and sort of ongoing feedback loops, but there’s a cost to learning, always, right? But when I send you a marketing message you may not like, the cost of learning is very minimal, so I will send you many of them, and I’ll do all kinds of A/B testing. But when I do that kind of A/B testing to groups or populations where the stakes are higher, then you might as well make that human judgment call of not doing it and relying on retrospective data.

NABA BANERJEE: Yeah.

SHERVIN KHODABANDEH: And I think you elucidated that point quite well for us.

NABA BANERJEE: Yeah. And one question that we ask ourselves is, what makes a human being trust another human being? What makes a human being trust Airbnb? And we are realizing that we can learn all kinds of models in the background, and you wouldn’t even know whether you were in
the risk threshold 1% or 2%, or marked as a good user.

But there is also a dialogue that needs to happen between Airbnb and you as a guest or you as a host, or between the host and the guest, that truly instills trust. Like when we ask someone, “Would you be comfortable sending your 17-year-old girl to a stranger’s house in Greece, when she wants to backpack before joining college?” You can probably tell this is a personal experience. I have a 17-year-old who’s about to go to college. And I realized that the first thought is, “No — not happening,” and I run trust and safety at Airbnb. But then when I think, “Well, if I could talk to the host; if she’s going to stay with someone else in a house; if that person is a mom like me ...”

SHERVIN KHODABANDEH: Yes.

NABA BANERJEE: And we as Airbnb could say, “We are running all these models in the background; don’t worry, your daughter is safe” — I don’t think that’s going to fly. I think as a parent, you will want to have the information you need to see, even though your judgment may not be better than the machine’s. So we are also working on, what is it that we need to say, that we need to do? And sometimes when we say things like, “Airbnb is not going to allow one-night stays that are booked at the last minute within the same neighborhood,” I think it makes a lot of parents feel better too, that OK, at least this option is gone. They will probably figure something else out. And that message cannot be convoluted. That message has to be simple. That message has to be clearly understood. And in addition to asking people what not to do, we also need to encourage people on what good behaviors they should do, like, talk to the host, talk to the guest, ask questions.

We have launched a program for solo female travelers because we were starting to see a slightly higher rate of maybe personal safety incidents in private rooms with solo female travelers. We started encouraging [women] find out, is there going to be a lock in your room? Is that lock going to be working? Will you have access to the bathroom just by yourself? Which spaces are shared? Which are not? So this was not a model that was stopping anything. This was simply an educational module that we had published in multiple languages, which was really helpful for our solo travelers. Part of my team’s work is not just to build invisible defenses but also to create trust and the perception of trust upfront through education and messaging.

SHERVIN KHODABANDEH: Very well said. I have to say, you mentioned you’re a mother of five. And I have to say, I think you stand on very solid ground when it comes to telling your kids what they can and cannot do, because you have so much data and so much experience, versus me when I say something and they’re like, “Well, how would you know?” And you could say, “Well, trust me; I do 70 million transactions a year, and I know what’s going on.”

NABA BANERJEE: And growing. Yes. I’m happy to talk to your kids, by the way. Anytime.

SHERVIN KHODABANDEH: Yes, that would be great. I’ll take you up on that.
SAM RANSBOTHAM: One thing about your measurements that I thought was interesting, maybe come back a little bit to that. You mentioned some positive metrics, too. So I think it’s a little tempting, and I think even in this conversation we’ve done that, is we titrate toward the negative. And that’s the news problem in general — is that bad news sells papers. Well, I think in the course of this conversation, we’ve shifted toward that. But some of the metrics you mentioned I thought had a more positive ring to them, didn’t they?

NABA BANERJEE: Yeah, absolutely. And this was an epiphany that we have been having over the past few years, because any kind of machine learning model typically does well when there is a lot of data to learn from. And with these kinds of cases, where out of 70 million-plus trips, when 0.1% of those result in even a report, and a fraction of a fraction of those results in an actual incident where someone gets removed, there is very little to learn from, and as a result, these models take time to mature. Of course, with technology evolving, that is going to get better, but we realized that we have a lot more data about good user behavior.

We have a lot more users who are actually coming in, booking their stay months in advance; they are probably checking in on time, leaving the property even better than they found it, communicating with the host and the guests, leaving honest reviews. So if we can actually flip the switch, and while we need to continue to look for the anomalies and the trends and the bad actors, if we can get really good at learning what good behavior looks like while making sure that we are also measuring for potential bias or discrimination and privacy compliance in how we collect and use this data, then that can be really powerful in informing when something does look risky. For example, if an account typically is accessed from the U.S., [and] suddenly we see that that IP is now accessing from the Philippines, and it feels like “Oh, this is an account takeover.” But if this is a good user who has typically been traveling around the world quite a bit, maybe this is not an account takeover, and there is history there from the good user behavior for us to be smarter about “What does normal look like and what does an anomaly look like?” Not broadly, but very specifically, based on the user segment.

So this is not new stuff, but I think we sometimes tend to obsess about just really learning what bad behavior looks like and getting good at detecting that, as opposed to complementing what an anomaly looks like [compared] to what normal behavior looks like.

SHERVIN KHODABANDEH: Yeah. And you have such a wealth of insights and information in a field that is really an evolving field. I mean, it’s not like a typical collaborative filtering situation where “users like you also bought this.” I mean, so much more goes on in terms of matching. This must be a very interesting and worthwhile problem to get your arms around.

NABA BANERJEE: Yeah. And it requires us to get out of our trust and safety silo and work really closely with our search, relevance, and personalization teams because our job
— unlike my job when I used to work in e-commerce — is not just to do the collaborative filtering and say, “People like you bought this, so you should buy this,” or, “You traveled to Paris, so maybe next you want to go to Rome.”

It’s about … you’re traveling with the family this time. You have little kids; you probably want daycare, and we know that this host offers daycare, and this location might be great for you. Or with the same customer maybe traveling alone, maybe we need to offer up different recommendations. And having these trust and safety signals embedded into our search-relevance algorithms can get really powerful in matching the right person with the right listing.

SAM RANSBOTHAM: Naba, you didn’t start at Airbnb. Tell us a bit about how you got there. You referenced Walmart and some background there, but how did you end up in your current role?

NABA BANERJEE: I was the first female engineer in my family. My father took a chance on me and said, “You’re curious. You’re a forever learner. Do you want to do engineering?” And I had no idea what engineering was at that time, but I signed up anyway and went on to become the first female engineer in my family.

I worked with Tata Consultancy Services, then went on to join Cognizant, came through Cognizant to the U.S., and worked for some time at AAA. I would just keep going to different projects and different jobs that would help me learn something completely new.

Walmart happened around 2006. I was asked to work on supply chain at Walmart, which is probably, if you want to learn supply chain at a company, that’s the company to go to. And while I was doing supply chain, every year I kept getting bigger and bigger responsibilities and projects until I found myself, from Walmart, leading almost all of the supply chain teams to going to Sam’s Club, which is a part of Walmart — leading all of product and front end and cart and checkout and marketing and pricing, which I had never done before. But it gave me a chance to get an well-rounded experience of learning how the back end of a large retailer works, as well as [how] front-end customer experience works.

And right after Sam’s Club, I was asked to lead search. Right around that time, I was starting to get really interested in a world that runs on machine learning and AI, and search was a great landing ground for me. I did a course from MIT, interestingly, to learn about the application of AI and ML, and that gave me the courage to take on the search job.

And from there, Airbnb happened, after 13 years, almost, at Walmart. I was so curious about the travel industry and this marketplace that was so different from anything I had done in Walmart. So the theme here is that I have always gravitated toward something that helps me leverage the skills that I have but then use my curiosity and learning to do something completely new and build myself up as I go along. That’s kind of my background.
SHERVIN KHODABANDEH: Naba, we have a section now where we ask you a series of rapid-fire questions, and just tell us the first thing that comes to your mind. What is your proudest AI/ML moment?

NABA BANERJEE: You’re asking me to choose between my different teams, which is career-limiting for me. But I will say that the work that we did on party detection and party risk reduction was groundbreaking, and there’s no one else in the industry who did it the way we did, so that is [a] really proud [moment for me]. But I am proud of all my teams, just for the record.

SHERVIN KHODABANDEH: That’s a great answer. What worries you about AI?

NABA BANERJEE: What worries me is how powerful it is. We had a chance to get to see [OpenAI CEO] Sam Altman up close and in person as he came to Airbnb. And just how fast this technology is improving and how much capability it has ... I worry about it falling into the wrong hands. I worry that while my team has this technology, the fraudsters have this technology too, and so I worry about, if today we are worried about fake IDs and fake spam messages being sent to our hosts and guests, how much more advanced this technology is going to get and how much more difficult it is going to get for us to detect the good from the bad. And I think it’ll be AI to the rescue as well in detecting fake AI.

SHERVIN KHODABANDEH: Your favorite activity that involves no technology?

NABA BANERJEE: Painting with watercolors. I love painting nature. And so paper, watercolors, and just peace and tranquility, no technology, is my favorite thing to do. And of course hugging my kids and spending time with them just listening to their day, trying to make them not be on their cellphones while they talk to me. Those two.

SHERVIN KHODABANDEH: Very well said.
The first career you wanted: What did you want to be when you grew up?

NABA BANERJEE: I wanted to be a teacher. I come from a family of teachers, and I feel so much joy when I’m seeing someone’s eyes light up with the gift of knowledge. I am a forever learner. Every job that I take has been completely different from my previous job; it’s because I love to learn. And so, yeah, that’s what I wanted to be.

SHERVIN KHODABANDEH: And you probably do a fair amount of teaching in your current role anyway.

NABA BANERJEE: I do a lot of mentoring.

SHERVIN KHODABANDEH: You’ve taught us quite a lot.

NABA BANERJEE: Thank you, thank you. I do do a lot of mentoring, because I feel like if someone else can see me do what I do, that will make them feel that they can do it too. So if I can just do that, that’s my life’s mission accomplished.

SHERVIN KHODABANDEH: What’s your greatest wish for AI in the future?

NABA BANERJEE: My greatest wish for the world, actually, is to not be so afraid; to give
it a chance. Because I think sometimes our fear of the bad holds us from embracing the good. There is so much wasted effort that goes into activities that should be automated through AI — so many patients who are not getting treatment; so many companies that probably need help and need so much funding to stand up basic things that can be done by AI; so many countries, probably, who are underdeveloped [but could] get so much advantage. I know that when it falls into the wrong hands, it can be used for bad, but the world has more good people than bad people, and I believe in the power of us using AI for good — using our collective goodness.

**SAM RANSBOTHAM:** I think everyone will probably resonate with your idea that the world is better than it is bad. We do tend to hear most of the negative stories, but it’s refreshing to hear how, one, how much attention you’re paying to try and to prevent those, but also to hear something about the good stories too that your platform enables. Thank you for taking the time to talk with us today. We’ve enjoyed it.

**NABA BANERJEE:** Thank you, Sam. Thank you, Shervin. My pleasure.

**SAM RANSBOTHAM:** Thanks for listening. On our next episode, Shervin and I speak with Elizabeth Anne Watkins, research scientist at Intel Labs. Please join us.

**ALLISON RYDER:** Thanks for listening to *Me, Myself, and AI*. We believe, like you, that the conversation about AI implementation doesn’t start and stop with this podcast. That’s why we’ve created a group on LinkedIn specifically for listeners like you. It’s called AI for Leaders, and if you join us, you can chat with show creators and hosts, ask your own questions, share your insights, and gain access to valuable resources about AI implementation from *MIT SMR* and BCG. You can access it by visiting mitsmr.com/AIforLeaders. We’ll put that link in the show notes, and we hope to see you there.