Me, Myself, and AI Podcast

Protecting Society from AI Harms: Amnesty International's Matt Mahmoudi and Damini Satija (Part 1)

SHERVIN KHODABANDEH: Many of our guests aim to use AI for good in their organizations. On today's episode, we speak with two researchers who focus on protecting human rights when artificial intelligence tools are used.

DAMINI SATIJA: I'm Damini Satija …

MATT MAHMOUDI: … and I'm Matt Mahmoudi from Amnesty International …

DAMINI SATIJA: … and you're 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 Khodabandeh, senior partner with BCG and one of the leaders of our AI business. Together, *MIT SMR* and 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.

Welcome. Today, Sam and I are excited to be talking with Matt Mahmoudi and Damini Satija from Amnesty International. Matt, Damini, thanks for joining us today. Let's get started. Matt, tell us a little bit about your role at Amnesty.

MATT MAHMOUDI: Absolutely. And, yeah, thanks so much for having us. I am an adviser and researcher on artificial intelligence and human rights at Amnesty's tech program. My role has been focusing on how certain AI technologies and, in particular, AI-driven surveillance are taken up by policing agencies [and] developed by companies, ostensibly for efficiency but often leading to discriminatory outcomes, inequalities of various forms, and affecting some of the most historically marginalized communities. So over the past couple of years in particular, I've been tracing facial recognition deployments, the companies involved, as well as where police are using the tools.

We've looked at facial recognition in places such as New York City, Hyderabad City in India, and the occupied Palestinian territories and [are] really paying attention to the ways in which these technologies that promised greater efficiency and promised to be sort of smarter ways of moving people from A to B or ensuring their safety are actually leading to the erosion of their rights.

SAM RANSBOTHAM: Matt, tell us a little bit about what Amnesty International does, what the structure is, how [its] technology practices got started.

MATT MAHMOUDI: Amnesty International is a movement of over 10 million people worldwide who work together via, for example, volunteering or through doing research and advocacy and campaigning to mobilize around key human rights issues of the day.

As far as the technology and human rights program is concerned, also known as Amnesty Tech, we're a collective of technologists, researchers, advocates, legal scholars, and more who work together on trying to hold both companies and states to account on their usage and development of technologies that really put those fundamental human rights at risk. So our work is to investigate and expose the ways in which those configurations of technologies are being used to erode those rights and, where possible, to advocate for stronger safeguards and regulations and human rights practices that enable us to enjoy those rights even as we continue to live through a rapidly changing world.

SHERVIN KHODABANDEH: Damini, tell us a little bit about what the Algorithmic Accountability Lab does.

DAMINI SATIJA: Yeah, thank you so much for having us here today. I work in the tech program, and I head up a team called the Algorithmic Accountability Lab, a relatively new team within Amnesty Tech. We look specifically at the increasing use of automation and AI technologies in the public sector and, within that, specifically in welfare contexts and social protection contexts. So we look at how governments and public-sector agencies are using automation to determine who gets access to basic essential services like housing, education, health care benefits, and so on. And our particular interest is in investigating and understanding how these tools have discriminatory impact or disproportionate impact on already marginalized groups, which is something we've already seen evidence of in public sector automation or automation of welfare.

And the team itself is a multidisciplinary team of seven individuals: data scientists, human rights researchers, advocacy, legal expertise — a whole range … to support the vision that we will take a holistic view in interrogating and understanding these systems' impacts on society.

SHERVIN KHODABANDEH: Thank you for that. This is quite interesting, Sam, because when most of our guests are talking to us about how they use AI, [it's] to create more profits or more revenues or reduce costs or do good, but generally, right? It seems like your role is to make sure we don't do bad stuff with AI, right?

And so in that context, given your background and expertise in AI, what do you think some of the guiding principles are, and how is it different? Like, when you're looking for bad actors, I have to imagine that it's fundamentally a bit different than looking to do good. I'm going to start with you, Matt. How do you go about doing this?

MATT MAHMOUDI: Well, oftentimes we learn about some cases involving a particular person that has faced a form of discrimination. In the context of New York City, for example, we were put in touch with an activist called Derrick Ingram, who has founded a collective known as the Warriors in the Garden but was also a prominent activist within the Black Lives Matter community. And he had been subject to harassment [by the police], who showed up at his doorstep and effectively harassed him for four hours for something that he didn't realize he'd done and, really, there was no clear answer to why they were there.

And as it turns out, given the presence of certain journalists around his home as he was being harassed, they figured out that the police had printed out a facial recognition identification report, which was present at the scene — which then, [it] turns out, had identified him as one of the only protestors identifiable at the particular protest, which was a Black Lives Matter protest protesting the murder of George Floyd. And in this context, we found out that, really, the police had simply identified this one prominent protestor with a megaphone and, as a result of being able to identify him, saw it as within their remit, even without a warrant, to show up at his door and try to question him and try to harass him.

The police eventually went on to come up with sort of a bogus charge in which they accused him of holding a megaphone too closely to an officer's ear, but this was all happening all the while Amnesty was investigating what other community members the NYPD had been targeting with this software and who was developing [and] providing the software that they were using. And the NYPD were not particularly forthcoming.

So our work has usually revolved around both conventional approaches, such as Freedom of Information Act requests or Freedom of Information Law requests, but it has also involved using, for example, Google Street View imagery to tag cameras that are run by the NYPD to give us a sense of how widely exposed New Yorkers, for example, are to network camera systems and, in particular, network camera systems with facial recognition. And that gives you a sense of how widely spread the risk is.

SAM RANSBOTHAM: So something that bothers me when people talk about artificial intelligence is this tendency, I think, to use anthropomorphic language. It's tempting to use phrases like "AI does X" or "AI does Y," and it's striking already [that] in talking to you both, neither of you have used *AI* as an actor. It's a tool, and you seem to be very focused on who's the actor. So the difficult thing there is that if a tool can amplify good and amplify bad, how do we promote a message to the actual actors? How do you get actors to use a tool that can be used for good and can be used for bad, to use it for good or bad? And even good or bad is tough to draw a line between.

DAMINI SATIJA: Yeah, and if I could use that to also piggyback to an earlier question where you asked about bad actors, I think that's very revealing in itself, because we are very focused on the actor, and it's not just the AI. It's also who has designed the AI, the way the AI's been designed, who's deploying it, what context it's being deployed in. And we have to be really careful not to focus on what is wrong with the AI because then that can also lead us down the trap of "There is a technical fix to this problem."

But often, the artificial intelligence tools we're looking at are also operationalizing a certain environment that we're concerned with, right? So, for instance, if we're looking at a tool being used in an immigration context, and the prevailing narrative is xenophobic or anti-immigrant, it will operationalize policies that fit into that category. So it very much is not just only about the technology, as you're saying, but also about the environment in which these are being developed, procured, and deployed.

And so that means that we're not always looking at bad actors as such, but bad use, to put it very simply. But I think that is just as much a guiding factor for us in looking for the cases we need to investigate, as is also, as Matt said, looking for discriminatory impact. I think another example that springs to mind here where a tool wasn't deployed specifically for negative consequence but it ended up having a negative consequence is a case of a housing algorithm that was used in San Francisco. And there was a story out on this a year or so ago.

There was a tool that was developed for social workers to use in allocating public housing. And the intent behind developing that tool was to provide something that allows social workers to have a more informed conversation with the individuals they're working with who need housing assistance. And the tool specifically would help them build a sort of vulnerability or risk assessment of the person to then determine how much housing assistance they needed. That tool was meant to help facilitate conversations. The way it was used, [however,] social workers were making yes and no decisions based on what the tool was spitting out on who should get housing assistance and who shouldn't.

So that … I mean you could argue that's bad use, but it's also kind of unintended use of the tool. So there are all kinds of realities that we're looking at that aren't as easy — it's just never easy to say that the issue is in the AI itself, which doesn't answer your original question but was some context that I wanted to add on the kind of bad actors question.

SHERVIN KHODABANDEH: It also highlights what you're saying — the criticality of AI and human [interaction], and not just one versus the other, or one or the other. Because in all of these examples, there are examples of unintended or unanticipated use, or maybe because of lack of training, or where the underlying narrative isn't that you start with intending to do harm; you just did not know or you did not anticipate that "Oh, I'm supposed to just use it as an input versus as an indication."

The one question I have … you alluded to it, but you went in a different direction than I thought you were going to go, because you said, "We're not talking about what's wrong with the technology, because the implication would be there's a technological fix." But I'd like to challenge that because why wouldn't part of the fix, at least, be technological?

DAMINI SATIJA: Yes, there are technical fixes when it comes to bias, and there are people out there who've put out ways of de-biasing tools. I think the reason we don't want to be completely confined to that is because of what I outlined — that we need to take a more holistic approach to understanding these technologies' impacts because, as we say, it's not only about the way the tool is designed, although, yes, that is really important as well. It's also about the human interaction with the tools and how humans use them.

And I think the other problem is that the technical-fix route can make us take a very siloed approach to what the problem is. So, for instance, in the AI ethics algorithmic fairness world, there've been a lot of de-biasing solutions put forward, and that implies that bias, in a very technical way within the algorithmic or AI system, is the only problem. But I think it's very possible that we could solve that from a technical perspective, but there are still myriad other problems with the tools that we're looking at. A, they can still be used in discriminatory ways, even if there's been a technical fix. There are surveillance concerns; these are data-intensive technologies.

We also worry often about sort of second- and third-order impacts of what these technologies incur. So, for instance, to take the housing example again, if a tool is used to deny someone housing or to deny someone access to Social Security benefits and then they're unable to pay rent or buy food for their family, those are effects that have happened two or three degrees of separation away from the tool, and it's still happening even if you reverse or take the algorithm out of the picture. That impact still exists and has still happened.

I think the emphasis, from our perspective, is maintaining that holistic understanding of the social consequences — political, economic — as well as technical. I don't know if Matt maybe wants to add anything on that.

MATT MAHMOUDI: I'd love to build on that a little bit further, in particular because of the housing example and other examples like it. Also, risk indicator algorithms that are used by children's protective services in order to make determinations about whether to remove a child from foster care or even put them in foster care. Especially work by Virginia Eubanks will outline how the social workers that are faced with this algorithm make determinations according to a light-based indicator that gives them sort of a red signal if it looks like there's been too many unsolicited reports of the child's welfare being in danger. And really, what that tells you is that the system in and of itself, the technology in and of itself, is not as easily fixed as, you know, to say, "Oh, well, then get rid of the indicators and turn them into more of a descriptive form of text." Because what you're dealing with is a technology that extends far beyond the actual code itself, which is what Damini is getting to here as well. It's an entire sociotechnical system.

You can't hold that AI is a thing without also holding that there is human-computer interaction that gives animation to how that system functions and what it does. So, what is written in the code — I've sort of taken the position — is somewhat irrelevant. What it does and what it ends up doing in the world, without using too much academic lingo here, but phenomenologically, is what really matters and what tells us about what the system actually is.

So by decentering ourselves from the notion that de-biasing is a virtue when it comes to AI technologies, and by decentering ourselves from the idea there's a technical fix from the system and instead holding that actually, these systems all ought to be tested and understood from what possible impacts they could have on society at large and on people's human rights before they're even entertained as being rolled out — that might lead us to the application and deployment of "better technologies." As for how we can use technologies to identify certain <em>harmful</em> technologies: An example that I brought before was how we were using street-mapping tools to get a sense of where cameras were. Just to be clear, we didn't use an image-recognition algorithm there. It was all people.

This allowed us to scale our volunteering efforts to some 7,500 people across the globe who helped us tag every intersection in New York City with cameras. That's a pretty, I think, compelling model for how you can scale activism and work that is moving the lever toward what might look like some form of justice and equity when it comes to technology, and certainly what an intervention that could promote greater respect for the right to protest might look like.

SHERVIN KHODABANDEH: This is, I think … My point was about technology. It wasn't to say, "Let technology fix the problem it's created," because the problem is created by the usage of it, as you said. And, of course, when you're talking about a powerful technology being used by institutions that have power to make policy, power to make law, power to make arrests or make war … of course the actor and the motivation of the actor and the use takes far more precedence [than] a technological fix. But I also have to believe that AI isn't going anywhere and the technology will only get improved.

And so I wonder … all of the deficiencies that your teams are finding, in terms of … I mean, in your example, you did not rely on image recognition to identify cameras, because you thought humans would be more accurate. Well, that is feedback to the algorithms and to the instrumentation that does image recognition. And in the example, Damini, that you talked about with housing, I wonder if there could be safeguards or additional prompts or additional data feeds that would actually make it almost impossible, for that technology that was making the choice on what to do, for a human to rely on the technological choice. So I must believe that as users and as agencies that are monitoring the use, that there is some feedback to the developer community that is building these tools. Not to say bias is the central problem, but that … I mean, you've highlighted so many different areas where technological artifacts could help advance the very cause that you’re talking about.

DAMINI SATIJA: Yeah. I mean, in terms of safeguards, there are many we could go into in terms of what we call for as a human rights community in regulation. I think Matt has already alluded to the No. 1 safeguard, which is clear questioning at the outset in the very conceptualization of these technologies as to whether they are required and whether automation is actually necessary in a certain context, and in doing that interrogating and scrutinizing, what that rights-violating or disproportionate impact could be of this technology. And I think in doing that, and what comes up for us again and again in our work is, which voices are being heard? Whose articulation of problems that need to be solved using technology are heard in that conceptualization phase?

And what we’re often up against in our work is that there are certain sets of pretty powerful voices, which you’ve just mentioned yourself as well. You know, policy makers, big technology companies, those who have funding to develop new technology, those who are funding new technology. Those who have the power to really dictate the trajectory of AI are the ones whose voices are also heard in what AI is being developed and then deployed, whereas those who are then impacted by the use of these systems, and especially the communities that we look at, often say … We’ve mentioned racialized impacts. Often, Black and Brown communities are really negatively impacted and harmed by these systems. Those are not the voices that are then feeding into what the problems are that need to be solved through this technology, which, as you say, is here and the development of AI is happening very quickly. But it’s that power imbalance that really concerns us in terms of whose voice is being heard what should be conceptualized. And that is an intangible safeguard but a very, very important one for us in our work.

SHERVIN KHODABANDEH: Very well said.

SAM RANSBOTHAM: It’s interesting you mentioned the social work example. My mother was a social worker and in [the] foster care [field]. And that is a highly understaffed, overworked world. And so when you gave that example, I have to say, part of me still finds it appealing that we could help those people improve. It may not perfectly correct, it may not perfectly do prediction, but given so much of what else goes on, it may be a better solution. So how do we get a better solution in place without opening up this Pandora’s box of difficulties to a point that we can improve it and can get experience over time? How does that happen?

MATT MAHMOUDI: So if I could jump in here, Sam, I think staying with the social worker example and just staying with a particular program that Virginia Eubanks looks at, it’s an interesting one because the state ends up spending more money on trying to hold up a failed technology than it would have spent just trying to equip the social workers with more resources, to be able to hire more social workers to be able to carry out their work more adequately and in line with the demand.

So I think, just drawing from the page of a piece of reading that I like to always assign to a class I'm teaching on science and technology studies, which is sort of a drawing from Chellis Glendinning's "Notes Toward a Neo-Luddite Manifesto," I will say I'm not anti-technology, and neo-Luddites aren't either, and I think that's kind of the crucial point here: that, A, neo-Luddites aren't anti-technology; they're worried about the ways in which technology creates numbers out of people and leads toward a hyper-rationality that takes out these important questions of harm.

And then, secondly — and this is a really important one — all technologies are political. We have to understand, what are the forms of politics and policies that are undergirding the particular deployment of a technology instead of, say, investing in the particular social programs that are required? So the kinds of examples that Damini has been bringing up all along and that we've been talking through really show that there is an insistence on investing in the tool of technology under the auspices that it's going to lead to some cost saving in the future, when the reality is that oftentimes states end up having to spend much more money either trying to hold the companies to account on what they promised but couldn't deliver or facing lawsuits by individuals, whether they're class-action suits or whatever, given the harms that they would've incurred on people, who have been subject to these mass forms of idealized technologies. Which I think goes to the point of, try and uncover what the politics that underlie it is and see if there is a social-political-economic fix that might actually be more sustainable than trying to get out of our way and go into this fantasy land of "AI will solve everything" — sort of technochauvinistic ideology that Meredith Broussard talks about — and get away from that a little bit, and thinking about what kinds of investment our society needs outside of these technologies.

I think, importantly, with tools such as your GPT-based chatbot models and what have you, you're dealing with systems that appear to be in perpetual beta, and so they can constantly make the claim that they're not working the way they should just quite [yet] and they may have unintended consequences because they haven't crunched enough data or quite gotten the model right. And you can sit on that narrative for a very, very, very long time.

But the question is, when do we, as a civil society, and when do we, as people who form a constituency on lawmakers that can speak on our behalf and regulate on our behalf, pump the brakes and say, "No, these are products that are out in the open. They're having an impact, and, therefore, they should be subject to regulation." It doesn't matter how large the language model is. It doesn't matter how much larger it needs to be to reach a saturation point at which it'll operate according to some prescribed fantasy of efficiency.

We have to get to a point — and that point, I think, was yesterday — in which we say, "We need regulation." I think [the European Unions's] AI Act, which Damini is working on extensively as well, is a really good first attempt at trying to create a regional-level legislation that has an understanding of the kinds of consequences we're dealing with and the kinds of impacts that these technologies can have on our rights and our ability to engage in the kinds of liberties that we have today.

SHERVIN KHODABANDEH: Damini, Matt, thank you so much for a very enlightening discussion.

DAMINI SATIJA: Thank you.

SHERVIN KHODABANDEH: Thanks for listening. Please join us next time, when we bring Matt and Damini back to continue the discussion about AI regulation, including what others can do to help limit harms stemming from the use of technology tools.

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.