Upol Ehsan Interview_final.mp3

Andrey [00:00:05] Hello and welcome to the 18th episode of The Gradient podcast,.

Andrey [00:00:10] The Gradient is a digital magazine that aims to be a place for discussion about research and trends in artificial intelligence and machine learning. We interview various people in AI such as engineers, researchers, artists and more. I'm your host Andrey Kurenkov.

Andrey [00:00:27] In this episode, I'm excited to be interviewing Upol Ehsan. Upol Cares about people first and technology second. He's a doctoral candidate in the School of Interactive Computing at Georgia Tech and an affiliate at the Data and Society Research Institute. Combining his expertize in AI and background in philosophy as work in explainable AI, or XAI, aims to foster a future where anyone, regardless of their background, can use AI powered technology with dignity. Putting the human first and focusing on how our values shape the use and abuse of technology, his work has coined the term human-centered explainable AI, which is subfield of expainable AI, and charted its visions.

Andrey [00:01:16] Actively publishing in top peer reviewed venues like CHI, his work has received multiple awards and has been covered in major media outlets. Bridging industry and academia, he serves on multiple program committees in HCI and AI conferences such as Neurips and DIS, and equally connects these communities. By promoting equity and ethics in AI he wants to ensure stakeholders who aren't at the table do not end up on the menu.

Andrey [00:01:46] Outside of research he is an advisor for Aalor Asha, an educational Institute he started for underprivileged children subjected to child labor.

Andrey [00:01:57] At Twitter you can follow him at @UpolEhsan - U P O L E H S A N.

Andrey [00:02:06] So I'm very excited for this, Upol has written to The Gradient before, and I think his work is super cool. Welcome to the podcast, Upol.

Upol [00:02:15] Thank you for having me on. It's pleasure to be here.

Andrey [00:02:19] Definitely. So as we usually do in these episodes before diving into your work a bit on your sort of background, I'm curious, how did you get into working on AI? I think your trajectory might be interesting or your background in philosophy as well.

Upol [00:02:36] Yes, I think I have Isaac Asimov to kind of attribute that credit to. When I was very young, I got hooked into his books. I have read forty seven of his books, not just the science fiction...

Andrey [00:02:51] Wow, that's a lot.

Upol [00:02:52] I mean, the maestro is is someone who's near and dear to my heart, which makes watching foundation and Apple TV right now a very scary prospect. Because I remember those things, but I think Asimov pushed me to think about artificial intelligence in ways that I don't think I would have thought of, because all of his books, if you think

about it, it's about how does how can we find flaws in the three laws of robotics that kind of he proposed, right?

Upol [00:03:26] And in college, I was very, so I grew up in a philosophy department that had a lot of cognitive scientists in them, but who were teaching analytic philosophy. And that's where I actually got into AI. I got hooked into it. I was like, OK, and maybe initially I had more ambitious goals of creating something like AGI, so to speak. But then over time, I started getting more practical about it. And after graduating, I actually spent a lot of time doing management, consulting and then ran a startup.

Upol [00:03:59] And in those experiences, I was dealing with AI-mediated applications, but mostly on the consumer side. So I had clients who are really using this at the enterprise level, and I was seeing how sometimes despite best intentions, the real use of these systems were suffering. So that's one way when I got into the Ph.D. journey, I started thinking of artificial intelligence, but from the human side.

Andrey [00:04:31] Right, and this was roughly when what year?

Upol [00:04:34] Yeah, so I had like so it was like I started the P.h.D journey roughly around 20 15/16. But the work that I had done before that was like the last four years before that, that's around like 2012 13. So that's like the industry experience very much drives a lot of my insights into the work today, especially seeing people and I do consult even now. So I'm very much in the applied setting of these research discussions, which help me kind of bridge too. That's why you'll see, even in my work, I do tend to have a more applied kind of a connotation.

Andrey [00:05:15] Yeah, yeah. I was just wondering because I think, you know, obviously there's been a huge boom in AI over the past decade and explainable AI, which you know your in has been more and more an area of study, but I think it took it a little while it sort of is catching up in some sense as as AI is getting deployed. Yeah. And then so you started your PhD journey in 2015, did you go to explainable AI right right away or did it sort of did you find your way there a bit later?

Upol [00:05:47] That's a really great question. No, I actually started my journey doing affective computing, so I was very much interested in helping children with autism, learn about non-verbal communication to head up displays, and Google Glass was very hot back then. Oh yeah. So I was trying to develop algorithms trying to help people who had had difficulties processing social signals to use some kind of a prosthetic to kind of augment that social interaction. So that's how I actually started.

Upol [00:06:20] And then after that, I am originally from Bangladesh. So I, the global south has been very much and is still very much a core part of my existence. So after that, I started looking at how do these technologies kind of behave in the global south, where the technology is not necessarily made in? After that, I think it was in two thousand sixteen or seventeen where DARPA had that XAI grant and that was the first time where because it's interesting, right? Like explainability of AI is not real. If you look at the literature in the 80s, there is a lot of work, in fact, that comics label and I was coined back in the 80s of the 90s. This was based on the knowledge, you know, the knowledge based systems like we had a second there.

Upol [00:07:13] But with the advent of Deep Learning and Deep Learning becoming kind of enterprise level almost like coming of age, you see, then there is this need to hold these systems accountable. So I actually had walked into my advisor's office at that time and I was asking, you know, what kind of projects do we have to work on? And he said, And my advisor is fantastic Mark Riedl. And Mark kind of said that, hey, there is this other project that no one really has taken upon themselves because we don't really know what it would look like. And I said, What is it this explainable AI? I think until at that time, like I had not heard about the term, I was like, This sounds like interesting. And I think upon reflection, what I realized about myself is I do very well when it's an empty slate and I get to paint my own picture rather than a very well formed slate. So I was very lucky to get into that debate very early on. In the second resurgence, I would argue, because the second life XAI has had is, I think, much more longer than the first life it had because it was there and but it also wasn't there in the early 1980s.

Upol [00:08:34] So then I started looking into it. I started on the algorithmic side, frankly, and then trying to work with algorithms. And then over time, I got on my Human side and you are right. I think explainable AI is very much in flux. That's how I would talk about it. I think we as a community, we are still trying to figure out how to navigate this field, being consistent in our terminology in the way we do our work. But there is also a certain level of beauty in that. And in that case, I'm kind of drawn by the social construction of technology lenses, something pioneered by way, a biker, and he talked about relevant social groups.

Upol [00:09:25] So in any piece of technology, you will have relevant social groups in that. Is why they was talking about bicycles, so bicycles have very other social groups, and each relevant social groups are these are stakeholders who have skin in the game actually give meaning to the technology as much as the technology gives meaning to the rights of. If you think about the mountain bikes and BMX bikes now, you know, like racing bikes on different bikes. And it's because of the stakeholders. They get very different, meaning all of them are bicycles, but they look very different. And I think within explainability we have people from the algorithmic side, basically the items from the HCI side. And now we are having stakeholders in the public policy side, in the regulation side, in the auditing side. So I think each of these stakeholders are also adding their own lenses to what is explainable and which is why you will see a lot of flux.

Andrey [00:10:25] Yeah, it's super interesting seeing this field kind of grow, and there's so much area to cover that. I think, you know, maybe compared to selling computer here. And you know, I think there's a lot more kind of maybe foundational or at least a conceptually important work. And then we'll get into what I think are yours and they could be could be called that. Yeah, your journey is really interesting. It's always fun to hear about how people bring in their experience before repeatedly and how that sort of guides their their direction. In my case, I started in robotics, in high school and then, you know, I did it in college. And then, you know, even when I went in some of the interactions and I came back to it. So it's always interesting to see how it happens.

Upol [00:11:16] I love that story because it's weird, right? Because I have an undergrad, I have a like a B.S. in electrical engineering and a B.A. in philosophy, right? And I never thought I would use that philosophy degree on a daily basis as much as I use it today. In fact, my edge in explainable air actually comes from my philosophy training because I can't access the writing that is coming from the air because even as academics are part of our training is how to read a certain body of work. But then when you're also trained in computer science, you can bridge it.

Upol [00:11:53] And I think there is something to be said there, especially for PhD student or other practitioners and researchers. Listening is I have been my mentors have always said like, you know, if you really want to make a name, pick an area and pick an Area B and then intersect them and you might actually get a C that is has a has an interesting angle to it that makes your work more relevant, more impactful. So I love also your story about robotics and how your back full circle. I think many of us in some ways are at the other end up where our interest kind of started.

Andrey [00:12:28] Yeah, for sure. It's it's quite interesting. You know, I worked in robotics a lot in undergrad and I worked a lot and then were kind of classical robotic algorithms not knowing you all nets. And then that definitely informed by understanding and my ability to get into it. So always, always call to see how that happens. So that's kind of my introduction to how you got here out of a way. Let's start diving into your work will be focusing a lot on a particular paper that I think is very cool. But before that, let's just give the listeners a bit of a conceptual kind of introduction to a field, I suppose, and then you general work. So just common basics, you know, quick introduction can you explain what explainability is, maybe, you know, and that's a pretty flat surface level and why it's important.

Upol [00:13:25] Yeah. So let's start with why it's important and then I'll share why what it is and I think the importance of drives what it is. So with with with today, like the AI powered decision making is everywhere from radiation radiologists using AI powered decision support systems to diagnose chest COVID pneumonia on chest x rays right to loan officers using algorithms to determine if you are loan worthy or not. Do you know the recidivism cases, right? So as we go on, more and more consequential decisions that we are making are either powered through AI or automated by.

Upol [00:14:10] So this actually creates a need for AI to be held accountable. Like if something is doing something consequential, I need to be able to ask why? Hmm. And the answer to that question is where explainable AI comes in. Broadly speaking, and many people have many different definitions of it, at least the way our lab and I have conceptualized it in the years of work we have done is explainable. AI refers to the techniques, the strategies, the philosophies that can help us as stakeholders within the AI system so it could be end users, developers, data scientists understand why the the system did what it did.

[00:15:04] And again, this is why it's also human centered in the sense that it's not just the algorithm, right? There's a human at the end of it trying to understand it so it can take many forms. Sometimes these explanations can be in the form of natural language, plain English, for instance, explanations like textual. Sometimes these explanations can be in the form of visualizations. Sometimes these explanations can be in the form of data structures, so they have the guts of a neural net where you are trying to figure out which layer is what's important. So these explanations and explainable AI I think the takeaway is very pluralistic. It's not monolithic. It's not. There's not one little thing that fits all. But at the core of it, it's about understanding the decision making in a way that makes sense to the user in a way that makes sense for the person interpreting it.

Andrey [00:16:01] That explains it. I think quite well. And I guess it's worth noting that this is especially difficult these days because we are working a lot to a Deep Learning. The way that works is you have a huge model of what awaits you. You trained on that on a

dataset. And then what you get is a phase where you can throw in an input and get an output right, and the challenges is, now explain why it's doing what it's doing, right?

Upol [00:16:27] Absolutely. Yeah. And actually, now that brings to another point, you know, there are many ways and then you hear different words being kind of used. In my view, I kind of split explainability into like transparency, interpretability kind of branches and then post hoc explainability. So I'll cover all eight of these. So transparency would be almost like clear boxing it so like instead of like black boxing it could you just make the model just completely transparent, like that's just one of the ideal to

Andrey [00:17:01] understand the model itself.

Upol [00:17:03] Then interpretability involves, I add in my view, the able to scrutinize an algorithm. So in other words, like like a decision tree life, like the infrastructure or the architecture of forms, the fact that I can poke and prod and I can get a good understanding and I can interpret what the model is doing right. But that also requires a level of expertize like you need to have the training to interpret a decision tree. You cannot just, you know, you can't just give anyone on the street like, Hey, here's a decision tree? Interpret it, right?

[00:17:35] So there's this level of interpretation that comes in, but the architecture of the model should also be able to support it. Not as you seem like deep learning algorithms are not really interpretive or by their architecture, right? Like they're not very friendly on their side. So recently, there has been a very big push towards what we call post hoc explanations. Right? So adding a layer, a model on top of the black box, so to speak, and make it somewhat transparent. So in other words, can I generate the explanation after the decision has been made? So those are the three main branches you see work within explainable AI these days, and a lot of people do use the word explainability and interpretability interchangeably. I don't. I tend to see explainability as a larger umbrella that can house, but doesn't mean I'm right to be honest, like it's being very precise about what you're saying when you're saying it. Does that help like kind of give the demarcation of the landscape as well the area in the work?

Andrey [00:18:38] Yeah, yeah. Of course it's it's interesting that at least you can think of it in these different dimensions, and I think that also helps understand sort of the ways you might approach it. And speaking of that, as you introduced in the intro, your work focuses in particular on human centered XAI, which is in some ways in contrast to algorithm centered XAI. So what is human centered XAI, in your view? Again, as kind of a surface level?

Upol [00:19:10] Yeah, it's about, I guess, the way to kind of think about Incentive XAI is the following like, there is a myth often in explainable AI, where we tend to think that if we could just open the black box, everything will be fine. Right? And my my my response to the myth is also. And not everything that matters actually is inside the box. Why? Because humans don't live inside the black box office, they're outside and around it.

Upol [00:19:40] And given, you know, humans are so instrumental in this ecosystem, right? It might not be a bad idea to start looking around the box to understand what are these value systems? What are people's ways of thinking that can ultimately aid that understanding ability that is so instrumental, explainable and so human centered, explainable AI? What it does is it fundamentally shifts the attention, and it doesn't say that algorithm centered work is bad by any means. It's not saying that what we're saying is we

need to put just as much attention on the human on who is opening the box as much as opening the box.

Andrey [00:20:27] Right? Do you need to sort of pay attention, care about the human aspect and not just think about the model, And then, you know, maybe we humans can take what you develop a model later and they can figure it out. That makes a lot of sense, and you have a great motivating example of this in your Gradient article having to do with this Fire Wall management thing and why human centered aspect was necessary. So, yeah, I find that very cool. Can you go ahead?

Upol [00:21:00] Yeah. So this was a this was a consulting project, but I had the privilege of kind of helping out with. They had a cybersecurity company, had hired me to address a very interesting issue of this firewall management system. And in that environment, one thing that happens is the problem was that bloat. So what is it? Bloat? Bloat is what happens when people open course on a firewall and forget to close them. So over time, you get a bunch of stuff that is open. But then what happens is at an enterprise scale, there is so many open course that is humanly impossible to go to every one of them and check. Oh, wow. Right. So they had a system that would analyze all these ports and suggest which ones do remain closed versus which ones do remain open. The problem was the problem here was rather tricky. The system was actually performing rather well around the 90 percent accuracy. It had really good algorithmic transparency. But the problem was, less than two percent of the workforce was actually engaging with it and using it.

Andrey [00:22:14] Yeah, and that's not what you want. Yeah, and then what was that? Yeah.

Upol [00:22:19] So and you know, I was brought in with the task of fixing this and the assumption was still back then and this was before we kind of coined the term Human-Centered XAI. And this is the project that actually drives a lot of that thinking. And the assumption was, you know, maybe the solution is within the algorithm, just fix the algorithm, maybe make it explain better, maybe open the box differently, so to speak. And what I found at the end of the day just to give a cut the long story short, I guess, is, there was nothing that was wrong with the algorithm.

Upol [00:22:56] The explainability that this company was looking for was at the intersection of the human and the machine not included in the machine. So what we found in this project presumption was still that something must be wrong with the algorithm. This was before we had coined the term Human-Centered XAI. A lot of the work here actually drove the philosophy behind it. And one thing that that came up was nothing was actually like we couldn't do much at the algorithmic level that helped the explainability of the system, the changes that had to be done, which actually I think we'll get into when we discuss the expanding explainability paper is at the social level.

Upol [00:23:42] So what was the problem here was people had no idea how to calibrate their trust on this system that without really understanding how others are also interacting with the system. Right. So for instance, if I'm faced with a new system and there is no notion of the ground truth, right? And the easiest example to share here was there was a young analyst and I'm using pseudonyms like Julie and Julie had a recommendation from the AI system and to close a few ports. And on paper, the recommendation was not wrong. I have suggested that, hey, you know, if you close these ports because they have been

open for a long time, they have not been used. So technically, these are not bad suggestions. Julie, not knowing a lot of the institutional history and how things are done accepted this decision.

Upol [00:24:38] Two weeks later, the company faced a breach. And then lost around $2 billion in one. What had happened was Julie had accidentally closed following the A's recommendation, the backup center reports. Right. So because their backups are reports, of course, it's good that they have not been used, right? It is also good that they're open. So this kind of highlights a very interesting tension here that even though the air system was technically not right, Julie actually got fired.

Upol [00:25:14] Oh, that's that's a shame. Yeah, yeah. So the accountability is squarely light on the human user, even though the human user in this case, they are not data scientist, either cybersecurity analysts, they shouldn't have to know how this guy is working. So it's very hard in real world situations to answer the following question one does this AI not know, right? And to address that question is almost an unknown, unknown, right?

Upol [00:25:44] You need and in this case, in this case study, they needed this thing. What the socio organizational context to help them understand how are other people dealing with it and and watching how others are acting with it, they were able to develop more robust mental models of how to calibrate that trust on the system. In other words, which are the situations that I want to see really well and which are the situations that AI does not perform really well because even if the performance is not uniform, that's the other reality in these real world systems.

Upol [00:26:17] So that's just, you know, just a quick summarization of this out of that case study, which kind of showed me that there were elements outside the black box that we really needed to incorporate in the decision making to help decision makers do it right and to make sure accountability was shared rather than be inappropriately placed all on the human and nothing on the machine.

Andrey [00:26:43] Yeah, yeah, it's interesting. I think a lot of listeners might now appreciate the importance of this kind of work in terms of, you know, outcome come here. And I think we'll dig in a bit more into where you want are in terms of how you do it, which was really interesting. Now, with a lot of these concepts laid out before we get into kind of our main focus, I thought it'd be fun to walk through kind of your journey in some sense of your trajectory, starting out less human centered and then sort of discovering that and more and more coming closer to where you are now. So first, you had kind of, let's say, a more traditional maybe XAI called rationalization and neural machine translation approach to generating natural language explanations. So just in brief, you know, what was this paper and sort of what was the contribution there?

Upol [00:27:46] No, thank you for asking that. I think this is the phase in my dissertation that I call turn to the machine. Mm hmm. I've kind of takes a few turns in this turn to the machine mark and I kind of end Brandt. So I just when I thought on my coauthors like Brant Harrison, who is at the University of Kentucky, Marl Riedl, obviously his tech and per now is also now I think is a PhD Student at Georgia Tech and Larry Chen, who is now graduated from Georgia Tech. We kind of started thinking that, you know, wouldn't it be nice if the AI system talk to you or thought out loud in plain English?

Upol [00:28:27] And the reason why we kind of thought about that was, Hey, I'm not everyone has the background to interpret models, right? And our a lot of our end users are not AI experts, but everyone, if they can speak and read and write in English, could understand English, right? In fact, that's how we even communicate. So I in this paper actually do a lot of inspiration from philosophy of language, namely the work of Jerry Fodor to kind of and work with Brant to kind of develop the algorithmic infrastructure to answer the following question.

Upol [00:29:05] And then this is the question that is asked me in this paper Can we? This is almost like an existence proof, like can we generate rationales from using a neural machine translation approach? And this was the first one. Yeah, to our knowledge, that uses an NMT mechanism. Instead of translating from like English to Bengali, like natural language to natural language, we we felt what if we replace one of the natural languages with some data structures? Right, right. And that's the insight in this case. And the innovation was we were able to back in the day, like when this paper was published back in 2017 18, there was a lot of work going on automated image. Captioning and stuff like that, but very little work was done on sequential decision making, right? So like, you know, if you can think of robotics, right, like getting a robot from one point in the kitchen to be in the kitchen is a sequential decision making task. So we actually took a sequential decision making environment and need an agent navigate it while being able to think out loud in plain English.

Andrey [00:30:15] If I remember correctly, this was like the game frogger?

Upol [00:30:18] Yes, yes. Yes. So that that was an homage to a lot of the game work that goes at the entertainment intelligent and Human-Centered AI Lab at Georgia Tech. So we kind of leveraged a lot of our game AI history, which I know, you know, I know you were at Georgia Tech for undergrad, so I think you might also be familiar with a bit of that.

Andrey [00:30:37] Oh, yeah, yeah, yeah. And yeah, Froger is a fine example because it's pretty intuitive, right? You know, why do you want to jump forward well as a car racing towards me so I wanna avoid it. Yeah, but that was a cool start and certainly interesting. But since when you have moved more towards the human-centered aspect, so that's going to the next step, I suppose return to the human, which I think started with this other paper automated rationale generation kind of extending this, but then a technique for explainable AI and its effects on human perception. So how did that come about?

Upol [00:31:17] So, yeah, so this one, so after we ask the question, can we generate? And the answer was yes. Now we ask the question, OK. These generated rationales, are they any good, right? Like because back then, if you think about how we used to evaluate these generative systems, you know, blue score or other procedural techniques are good, but we don't really get a sense of how good they are to human beings. Right? Like, do people actually find these plausible?

Upol [00:31:48] So in this paper, ours are like kind of starts to turn to the human. We presented the first work that gave a robust human centered use our study along certain dimensions of user perceptions to evaluate these rationale generating systems. And what we found was that we bridged a lot of work. So I took all of these measures and adapted it from work in HCI human robot interaction, as well as the technology acceptance models from back in the 90s when automation was becoming hot.

Upol [00:32:25] And we found fascinating things around, not just the fact that these were plausible in this paper. We just didn't make the Frogger kind of say things. In one way, we were able to tweak the network in a way that I could make Frogger talk more in detail versus, say things more shortly in its rationales. And we found that the level of detail also had a lot of interesting interweaving effects on people's trust, people's confidence, how tolerant, where they when the robots like that further failed, right? So this was a really interesting deep dive.

Upol [00:33:05] And we just not only did the quality quantitative part, we did a really good qualitative part as well. These are the crowd workers. And, you know, getting Amazon Mechanical Turk first to take a forty five minute task is not easy, I think. And so we were liking the methodology part. I think we were very happy with it, and I'm so proud of the team that did it. They were saying Han and another research assistant were undergrads at Georgia Tech who helped us create a really good data collection pipeline that helped us collect these rationales to train. And then we not only train, but we also tested it. So that was the end to end kind of application of this that really made the paper one of my favorite papers that I've written.

Andrey [00:33:52] Yeah, is this reminds me a little bit of the whole like. Some field of social robotics is quite interesting because again, there's a lot to do with human perceptions and like, how do you communicate intent of grasping a couch in a way that you know, people can understand? Or how do you appear friendly and so on? That's its own whole thing, and it's always interesting to see that, you know, aside from all of the social models, if you need air during the real world, this is also a big challenge indeed.

Upol [00:34:28] Yes, so in this one, one aspect that differs from the other white we'll we'll get to soon is here, you sort of are still dealing with one to one interaction versus playing game and then the agent is kind of trying to make it clear what's going on. And you already mentioned in your example that you know you need and many real war situations to go beyond that, you need organizational context. You need to understand groups of people, so to speak. And that takes us to the concept of socio technical challenges. Yes. So how did you make that turn and what is that compared to this one to one paradigm?

Upol [00:35:15] Absolutely. So you hit the nail on the head, right? There is like a lot of the way we were thinking about the rational generation or the interaction paradigm was very much one to one. And, you know, based on my prior work in industry settings, I started realizing that is that truly representative of what happens. And I started realizing that, no, we need to think more about like these AI systems. I'm never in a vacuum. They're often situated in larger organizational environments. So in that case, how do we think about this? How do we conceptualize this?

Upol [00:35:52] So this kind of forced us, and this is probably the first kind of conceptual paper that I have written is to kind of outline. So we kind of coined the term human centered essay, but we also wanted to seen how do you operationalize this thing? So we bridged theories from critical AI studies like critical technical practice in HCI, like reflective design and value sensitive design. And we kind of talked a little bit about, OK, now we have this insight that we have to not just care about one person, but also multiple stakeholders in the system. So going back to the cybersecurity example, right, it's not just the analyst who is making the decision, it's also the decision of the analysts previous who had made similar decisions in the past. So that kind of forced us to kind of imagine and

envision AI explainable AI paradigm that is more human centered and not just one human, but also incorporates many humans.

Andrey [00:36:52] Mm hmm. Yeah, so there comes a socio technical aspect. You know, social being, you know, interactions between people and even organizations. So where you marry sort of the groups of people with the technical problem, which now you really need to think about both. And that as far was sort of kind of new direction that wasn't really the norm or stylish in the field.

Upol [00:37:23] Yeah. And I think that's a very important point. In this case. I had drawn a lot of inspiration from the fact literature of the fairness, accountability and transparency literature where they were very much at that time thinking very socially or technically. And I am always reminded of I watched this video from Microsoft Research is like responsible. I kind of in visions. And Hannah Wallach, who is at Amazon New York, had this fascinating line that I cannot like repeat verbatim. But the version that I remember is today. Our systems are AI systems are embedded in very complex social environments.

Upol [00:38:05] So that means out the effects that these technical systems have our social. So that means that fundamentally socio technical in nature, in terms of their complexities as well as that impacts. So when we keep that in mind, I started asking myself, how can we get a good idea about explainable AI if we do not take a socio technical perspective given, you know, in the real world, that's how these systems are. So that's actually a lot of the things that drove these socio technical lens, so to speak. And you are right, like this was the first paper, to our knowledge, to kind of highlight that explicitly in the context of explainable.

Andrey [00:38:49] I yeah, I find it interesting. I think it it seems like it would be easy to not have this realization if you come from a traditional sort of AI computer science research background where you just jump into a Ph.D., you know where you work, in your office, in the computer science building, you know, doing your research. And it's easy to forget sort of about the outside world. So I think it's interesting also that having had all this background outside working in actual organizations, I think I would imagine that also made it easier for you to get here.

Upol [00:39:28] Yeah, it was. And it's humbling, right? Because you fail so many times trying to do this, and that's the only way sometimes we learn. Right? I, you know, my consulting projects, I never like linear or straightforward because they often reach out to me when problems are so complicated that in-house teams need external help. And I think, you know, a lot of us then learn the lesson that I have learned through all of this is embracing a sense of, you know, taking a learning mentality from a lot of the there is a famous paper. I forget the name of the author who kind of framed mistakes as mistakes like, you know, in a movie, you take multiple takes and not all the takes work. So a lot of them are mistakes, right? So I really embrace that mentality of mistakes. Not all projects will work out. You have a lot of mistakes, I guess. Nothing is a mistake, per se. And I think that really helped me have a more iterative mindset, which has paid a lot of dividends in getting a lot of this work done.

Andrey [00:40:31] Yeah, I think it's interesting how in some sense, especially doing it really enforces that. You really have to adapt to that because you got to have failures, but almost always will inform your understanding and ultimately guide you to something interesting, ideally, you know. Yeah.

Andrey [00:40:51] So as you did research that led to this paper, human centered explainable AI towards a reflective socio technical approach where you lay a lot of groundwork for how you can move towards that. And we we really can't get too much into it. It's it's quite detailed and self. But he did write this excellent piece on a gradient towards human centered, explainable AI every journey so far. So we're going to link to that in the description, and you can just fly a gradient and recommend you read that. But for now, we're going to actually focus on a more recent work expanding explainability towards social transparency in AI systems. So to get into it before even getting to any of the details, you know, what was your goal in starting this project and sort of a problem that motivated it?

Upol [00:41:51] This is. Frankly, I feel like this was the paper that, like the Human-Centered AI paper, was the paper that needed to be written first for me to actually write this paper. And you know, a lot of the work in that cybersecurity company kind of really informed this. So, you know, for the longest time, I have been kind of arguing that we need to look outside the box, right? Yeah. So then, you know, largely speaking, the community will come back and ask her to call. I kind of get what you're trying to say, but what outside? What is outside? What do you want us to think about? And the the kernel of this paper is fundamentally and as as the title kind of says, extending explainability, it expands our conception of explainable AI beyond the realms of algorithmic transparency.

Upol [00:42:43] By doing what? By adding this new concept called social transparency, which is actually like, not like New New in the sense that we created it in the in the context of XXXII, it's new. There is sort of transparency in online systems back from the 90s. And in the paper, we kind of pay homage to a lot of those work, but it's fundamentally making the following observation. So within AI systems, and I think this is where it becomes very tricky when when we say AI systems is actually somewhat of a misnomer because when we say AI systems are a very important part is left out and it's often implicit, which is the human part.

Upol [00:43:24] Implicit in AI systems are what we call human AI assemblages. Right? So these are two coupled points. So ideally, what you're really going for is the explainability of this assemblage, right? The human part of being often implicit. But but how can you get the explainability? All of these assemblage, these two part system, the human and the AI by just focusing on the air and asking the question that is asking in this paper? So then the question becomes IRA to Paul. I get it like, you know, you can add, know you need the human part, but what about it in looking very Typekit? So that's where if you add the transparency on the human side, we kind of introduce this notion of social transparency in AI systems, right? Operationalize a little bit of this in the paper.

Andrey [00:44:14] Right. So yeah, it's I think very interesting about this in terms of operationalizing that. Not only do you highlight this need, which I think is very intuitive, but actually exquisite talk about how to do this, how to be useful and really beyond technical transparency and how to integrate that. And actually, you know, where do you start? How do you how do you do it and so on, right? Hmm. So I guess maybe can we dove in a bit more about this idea of social transparency? Ethics is so important. And so we know because fancy sort of trying to understand what the algorithmic side of it is doing, what the model is thinking, but what is social rationality one of its components? And yeah, how should people understand it?

Upol [00:45:06] Yeah. Well, that's a that's a fascinating question. So to understand social transparency, I think we have to accept a few things. First, we have to understand and acknowledge that work is social, right? You know, we don't work in silos. Most of us, we work within teams. So there meet. That means right. There is some need to add this transparency data in an office when you're working with a team or virtually through Slack, right? There's a lot of chatter that is going on and there is a necessity behind that. So that is fast the the fast realization that work is social. That means there might be a need to make that social nature a little bit transparent, especially when we're dealing with AI mediated decision support systems. So and you know, as we share in the paper, we were trying to.

Upol [00:45:59] So this is also difficult as well to some extent, right? One of the challenges that AI researchers face is how do we know what the future really looks like without really investing months and months of work, building large infrastructures and models and then realizing we're actually not very useful, but that that is a very hard cost of doing this. So due to kind of explain that we used this notion of scenario based design. So this is coming from the traditions of design fiction, as I'm drawing a lot of this actually from the theoretical underpinnings of the human centered explainable AI paper that we just talked about. Mm-Hmm. And so using scenario based design, we conducted around four workshops with a lot of people. From different technology companies, just to get a sense of what are the things that are outside the black box that people want when they make a decision within the AI system. Right. So that's the workshop is meant to kind of get a more formative understanding, right?

Andrey [00:47:09] What needs to be made transparent in this social system in terms of

Upol [00:47:13] right, because there are so many things you can make transparent, right? Because how do you know which one is the right thing to do right? And I think through these workshops and this is pre-COVID, so we have the ability to kind of get in person and kind of have these workshops. And what we learn was that out of this and this is, I think, what we what we call in the paper, the four ws, right? So in addition to the i's technical transparency or algorithmic transparency, these practitioners, data scientists, analysts and others wanted to know for things who did what, when and why.

Upol [00:47:58] So those became again, we are not saying this is the end all, be all to all social transparency. There might be other socially transparent systems that that do very well, but actually in the cybersecurity example, going back to that when we implemented this aspect of who did what, when and why. So imagine, like, you know, next to a threat, you know, close these ports, right? Let's imagine that if Julie have social transparency, what would do we have seen? Julie, who got fired before? So when you get this new disease, the air is recommending the ports to be closed and you're like, OK. Is that true? Like, is that real or not? I don't know if it's a false positive. But then Julie is able to see, you know, maybe 10 other people dealing with a very similar situation in the past. And in one of those Julie scenes, I one of the who's right? Maybe imagine this is Bob, and Bob is a veteran in the industry. He's like a level three analyst, and he says, Oh, these are backup site reports in law. Mm. Right? So who did what? Right? When? Maybe, let's say, three months ago? And why? So the why is the reasoning right? Like these are backup center reports ignored by situating this extra piece of information that is actually capturing? You know, one might argue, Hey, people like that seems like a bad problem. They should've just added back to the data center, right? That's not good. And that is where I think the

critical insight lies. There is not enough things you can add in the dataset. It's like a golden goose chase

Andrey [00:49:37] because it's all inside the model, right?

Upol [00:49:39] Exactly. And sometimes things happen dynamically. Remember, data sets are basically snapshots of the past. And work norms actually change over time due to the sensitive nature of certain cybersecurity explanation institutions. You do not want certain things to be coded into a data set. Right. Because what if that gets hacked, then all your secrets are out. So there will always be elements that are not quantifiable, that are not acceptable in a cleanly named dataset. In those cases, those very things that are hard to quantify, hard to incorporate often can be the difference maker between right and wrong decisions where they are. So by adding this social transparency, you are able to inform someone to know when to trust the AI versus not.

Andrey [00:50:33] Mm hmm. Yeah, exactly, and to dig a bit deeper. I would love to hear how did this scenario based design process work? I think figure one of your work is really interesting is that the scenario used.

Upol [00:50:50] So you know, in this scenario, we asked our participants to kind of envision being in using a AI powered pricing tool to price and access management product to a customer called Scout. Right. So the AI kind of does its analysis and recommends that, hey, you got to sell it at 100 bucks per month per account. And it did also share some post-rock explanations and kind of justifies, White said, what it's saying and that the model, the technical transparency pieces like the item, the quotable goes off a salesperson into account. It did a comparative pricing of what similar customers pray, and also it gave you the floor. So what is the cost price for doing this product? So these are so imagine that's the first letter and today, right? That's the state of the art. Nothing is better than that, right? We don't have the social transparency that we kind of envision in this paper, but that is where the state of the art was.

Upol [00:51:47] So that was our grounding moment. So we would ask our people as they went through the walk to what would you do right now? Do you think this is, you know, before we showed them any social transparency? Right? And we will see that most people agree with that. Yeah, this seems like a decent. We also kind of calibrated the price point by asking experts. So we kind of grounded a lot of his data, even though it's a scenario. If it's fictional, the fiction is grounded in reality, our version of reality. And then we told them, like, now imagine what have you found out? And that only one out of 10 people sold this product and the recommended price? What would you do?

Upol [00:52:24] And you could see our participants kind of get very interested, like those like, oh, that's really interesting information that helps me calibrate what to do. So then we kind of dug deeper, which is like the bullet points three four five to share three examples of past colleagues who have dealt with the same customer scout on similar products. And then one of the most important comments were made by Jessica or Jess, who's a sales director. And it turns out Jess had rejected the recommendation, but the sale did happen, and the comment was the most important where they say that, Hey, is COVID 19. And this was done at the height of the pandemic. I can't lose a long term, profitable customer. So they offered 10 percent below the cost price. And that's an important part, right? That not only did they give you a discount, but just the director had given them below the cost price and that social context of what was going on that was outside of the algorithm, right? Very

much inform how people acted on it because remember, without any of this context, they fell. The price was fair. It was done the right way. You know, the justifications were right, but very few people actually, you know, offered the same price when they knew what others had done, especially when a director level person had done it before.

Andrey [00:53:52] Yes. So in a sense, I think going back to something you mentioned, it's letting you know what the model doesn't know. Right? It doesn't know about COVID and these four W's. The social scenario doesn't really explain the mottoes decisions, but it does like to understand via a system better in the sense of like the AI system is situated within the organization. And so you get to know more its weaknesses and where and when to follow it. Maybe when not, which I think would be a lot harder about seeing like, OK, this first person accepted discrimination. This person didn't. And this figure, I think, illustrates that really well.

Upol [00:54:36] And I think it's kind of asking the question like, what are the eyes blind spots and can other humans who have interacted with this system in the past and address it? So like, for instance, I am now currently working with radiation oncologists on a very similar project and in radiation oncology, just like in other fields, there is no absolute ground truth. With 80 senators, we are so comfortable with the terminology of ground truth, right? But when it comes to using radiation to treat cancers, they're established practices. There is no like absolute gold thing that everyone must do because each patient is different. Each treatment facility is different. So in that case? Knowing when to trust these recommendations, foreign by saying, you know, give this much radiation to the patient's left optic nerve. Right. That's a very high stakes decision, right? Because if you do it the wrong way, you can blast away my left optic there and take away my vision. Right? But guess what? What the AI system might not have known is that the patient is blind in the right die. So all of the calculus goes away because we know there's no like central blindness data in randomized controlled trials. Right?

Upol [00:55:52] So just knowing that extra piece can help you calibrate how much treatment you want to give it and knowing what your peers done right, because in these kind of communities of practices is very much community driven, right? Like the radiation oncologist kind of have these standards that they co-developed together are two studies. So this social transparency starts mattering extremely when the cost of failure is also very high, right? Like, you know, blasting someone's optic nerve nerve out there, the radiation is a pretty high cost rather than, you know, missing a song recommendation. And I think that's the other part like you don't think I don't. Social transparency is not really helpful when the stakes are low or when the nature of the job is not very collaborative, right? But the more the stakes are high, the more collaboration is needed. Social transparency becomes important because what then becomes very social.

Andrey [00:56:49] Yeah, it's just makes me feel like you can almost consider this like what if the AI model is in some sense, a coworker, right? When you work with people, some people you trust more and less and when you do decision making, it is sort of collaborative. You know, you might debate, you might ask whatever you consider, address, if you consider that that's not something that you can do with any AI system, at least for now, you can't say, well, you know, have you taken this into that into account? But seeing the social context, it seems to me, was what people might have realized that to me, this comment and now, you know, didn't take into account the COVID thing. So, yeah, she gets interesting in the sense of like you get to know the system as another entity you work with almost.

Upol [00:57:43] Yeah, yeah, exactly. And I think that's kind of changes the way we think of human collaboration, right? Because you are now like, I often think about it, it's like, you know, Avatar The Last Airbender. I don't know.

Andrey [00:57:55] Yeah, yeah.

Upol [00:57:55] So it's like, what does Avatar do? And new faces around there, like Avatar and right? Like when he faces some difficult choices, he kind of seeks the counsel of past avatars who had come before him. And so the social transparency is in a weird way of capturing that historical context in a system in situ that really makes your decision making in that moment much more informed. Because at the end of the day, we have to ask ourselves why these explanations aren't there. They're there to make things actionable. If you if someone cannot do something without explanation, then there might not be any explanation, right? Like if the machine is explaining itself and I cannot do anything with it, that's very difficult. Like, I don't know what purpose of serving other than just understanding. But if I understand and cannot do anything with understanding, what is it there for anyway?

Upol [00:58:50] So social transparency can make things more actionable, even if right, even if. The participants are saying no to the system, and that's the crucial part. I think it changes how we formulate trust because a lot of the work around trust that we see is around user acceptance. I want my user to like the I want my user to accept me. But I think what we are seeing is it's not about just mindlessly fostering trust. It's about mindfully calibrating trust. You don't want people to over trust your system because then there are liability issues.

Andrey [00:59:30] Yeah, exactly. And, yeah, so in terms of this process, you started these four workshops you, I think, carried out this idea of the four W's what who why when. And if I understand correctly after the workshops, you then had sort of a more controlled study where 29 participants. Is that right?

Upol [00:59:53] Yeah, yeah. So then then we really did this once we built the scenario, right? Then you kind of when made people go through the scenario of the study. So we would walk them through the scenario just the way I kind of described a few minutes ago. And we will start seeing that how they start thinking through this and this is the beauty of scenario, these design, right? You can think of this as a probe. So you are probing for reactions about an air power system without really needing to invest the severe infrastructure that is needed to make a like a full fledged AI system. But you're getting very good design elements out of this for a lot less cost. It does not mean that you don't build a system, of course you do. So, for instance, in the cybersecurity case, once we did very similar studies with them, with scenarios, we went and we built out, this takes. And guess what, right? Like two years into the project with them, that company actually now lives in a socially transparent world where all their decisions are actually automatically situated with prior history. And they are actually you able to use the four W's as training to retrain their models so that the decision is not only just algorithmically situated but also socially situated, right?

Andrey [01:01:14] So it becomes of this sort of feature space to model is informed by it seems.

Upol [01:01:19] Absolutely. And then think about it that way, right? Like you are also getting a corpus without actually building a corpus. Right. Because over time, what is going to happen? These four W's are going to get enough density depending on who knows, right where they become large enough that you can feed back into the model. But the cool part is from day one, they're giving value to the user. Right? So they're not like being a grunt work of a data set building task. That is the one thing that a lot of my cybersecurity analyst stakeholders would counter that like, Hey, this is actually useful. I like doing this because it doesn't make me feel like I'm building a stupid dataset that I might not ever see. So they're actually building a corpus while getting value from it, which is very hard to achieve in any kind of dataset building tasks.

Andrey [01:02:09] Mm hmm. Yeah, exactly. And it's interesting to hear that they are more into it. And I think it's also funny that, you know, having done a study in the paper, you are able to not just give you on tape, but actually quotes the study participants and really, you know, showing their words very concretely how you came to your conclusion. So for instance, one quote that I think is really relevant is I hate how it just gives me a confidence level in gibberish to engineers will understand for zero context, right? It's a very human reaction that really tells you, Well, you know, this person, what's the context, right? So then, you know, we've talked through somebody resembles your study, and now I think we can dove in a little more. So we've talked about social concerns. And I think in the paper, you also break down a little bit what exactly is made visible, what you want to make visible. So the decision making context and the organizational context. So yeah, what is involved in these things that people really need to understand?

Upol [01:03:23] Yeah. So, you know, through the four W's. So these are the kind of like that you can think of them, the vehicles that carry this context, right? Mm hmm. So the four is the first thing, and I think we kind of shared a framework that sees how it makes context visible at three levels, which do the technical, the decision making and the organizational right. So but with the umbrella of all three is the first to use what we call crew knowledge, right? So crew knowledge is really an important part of these informal knowledge that is acquired through hands on experience. It's part, and it's tacit of every job that anyone has ever done. Right? And it's often situated locally in a in a tight knit community of practice, sort of like an aggregated set of Milhouse. Right? So the Y right, the Y is actually giving insight into that knowledge.

Upol [01:04:24] These are the variables that would be important for decision making, but sometimes are not captured in the eyes kind of feature space. Right? The other part is like social transparency can support analogical reasoning in terms of like, OK, if someone has made the decision in the past, like you remember, just I just gave a discount for the COVID case. So that means I too can give the discount on the Kovic case. Right? So it's aiming. So at the at the technical level, it helps you calibrate the trust on the air right and the decision making level. It can foster a sense of confidence and a decision making resilience that you know. How good are you? Can you trust AI? Can you not trust and self confidence, right? Yes, these difference between like, do I trust the AI versus do I trust myself to act on that? And I think those two are slightly different constructs and organizationally leases.

Upol [01:05:22] So for them to use is capturing these tacit knowledge and the meta knowledge of how an organization will work. So you get an understanding of norms and values, right? It kind of encodes a level of institutional memory, and it promotes accountability because you can audit it, right? If you know who did what, when and why

you can go back and audit things. So those are some of the things that we got out of this that are helpful when it comes to making the AI powered decision making.

Andrey [01:05:53] Yeah, it's quite interesting this notion of decision making and organizational context. I think you define decision as sort of, you know, localized to a decision. So like, you know, you choose a price quarter, you you think about similar price quotas. Organization context is something that's easier to forget, but it's sort of, you know, what do we stand for? You know, how aggressive are we? You know, these sorts of things that are or in general and yeah, pretty interesting to. I think and I guess you came to understanding this sort of split by just seeing what people used or included in their four W's.

Upol [01:06:37] Yes, I think this came back from those workshops. I think where we kind of understand like, OK, because, you know, we were thinking, maybe there is an h like how maybe there is another W like where. So we were trying to understand what would be the minimum viable product, so to speak about in the in the social transparency, because we didn't also want to overwhelm people. So the way we kind of understood like what they were doing is actually through the case studies that we have done in addition to this study. Right. So we were trying. We were inspired by that aspect. And that's why, like, you know, in the in table three, we kind of talk about, you know what? What was it? So it's an action taken on the API, the decision outcome. Why is like the comments with the rationale that justifies the decision because of what and why is always linked? And then who the name the organizational role, because sometimes seniority starts playing a role like it was a director. You kind of take their their view a little more than others. And then when is the time of decision? And that's important because sometimes something some decisions are not relevant, right? So think about like, you know, pre-COVID, decisions do not become very relevant during COVID. So those are some of the things that we found, not just by our workshops, but also analyzing the data, the qualitative data through the interviews and the walk throughs that we had.

Andrey [01:08:05] Yeah. And then you found, you know, the what is really important, the why is important, the when, you know, sometimes, but you know. Yeah. And then also, I think probably informed the UI, sort of how you present things.

Upol [01:08:19] We actually asked our participants at the end of it. I'm like, Can you rank it and tell me why? Right? So we would make them rank the them like, tell me what you cannot live without and everyone saying, I can't live without the what. And then I said, OK, imagine that I can give you one more. What would that be like? Oh, I need another wide. And then I said, Now imagine I give you one more. That's I need to know the whole. So that's how we kind of made them do this ranking task and then get a sense of importance, because sometimes many companies might not have all before that. There might be privacy concerns that prevent the HU from being shot. Right. Because you can also see, like, you know, biases creep up, like if I show the profile picture and you know, if you can guess the person's race or gender from the profile picture, it can create certain biased viewpoints or even the location, right? Because certain companies are multinational and it could be that, you know, certain locations are not often looked positively enough. And that's my bias, the receiver's perception. Mm hmm.

Andrey [01:09:23] Yeah, exactly. And I think again, it's interesting here, just reading the paper, which I think is, is, you know, I would recommend it. I think it's quite approachable. Is again, you have these quotes from the study participants that make it very concrete.

One of them is the outcome should be a tldr the why is there if I'm interested, then there's also the issue. Someone said if I knew for to reach out to, I could find out the rest of the story and so on. So again, it's it's really giving you a sense of how your study and interaction with people led you to your conclusions, which I really enjoyed in reading the paper.

Upol [01:10:08] No, you know, thank you so much for the kind words we put a lot of love into this paper.

Andrey [01:10:14] Yeah. And, yeah, you know, that's what you need to make a paper really enjoyable, so your work pay off. Now I think we can touch on a lot of it and it went through, I think hopefully the most this stuff in terms of the study elements and the four W's and make clear a social transparency is now on to a couple final things. So we've said, you know, it's good to have this on top of what is already there. I'll go to make sure it's fancy. So you need to add the social transparency and one question there as well is it easy or is it a very challenge is in place that would make it harder to do that?

Upol [01:11:00] Yeah, that's a that's a good point. I think, you know, as as with everything, there has to be the infrastructure that is supported, right? And there are challenges like privacy. There are challenges like biases or information overload, as well as incentives like if you want to engage in a socially transparent system that has to be incentive for people to engage with it like, you know, give those four W's as they're working, that is a burden that is added, right? Like no fees, lunches. And so that means we have to be very mindful of that. And you know, we can, you know, with the Ford family, you could also kind of promote groupthink, right? Imagine in a company culture where you're not allowed to go against your boss and you see a comment from your boss previously. So so we have to be careful. You know, it's not a golden bullet. So we have to be very careful when we operationalize the social transparency that we are trying to be very mindful of some of these challenges, like, you know, do we really want to see all the four W's at every single time? No, there are ways to summarize it. And we have done that in my project with the cyber security people, we have been able to figure out how to summarize these aspects at a level of detail that is actionable.

Andrey [01:12:17] Yeah. And so speaking of cyber security, people say, take it outside the study, I was interacting with these participants and, you know, figuring out the of context. You also took this for context to an actual organization and then tried it out.

Upol [01:12:38] Is that right? Yeah. So like if you remember just from a timeline perspective, right? So by the time I think we wrote the paper we already had. This is obviously the study. So there was an empirical study that was done. Separate from this in parallel was the cyber security project that I was running for a long, long time. And what I had the, I guess, the luxury of knowing the future to some extent is we were able to incorporate a lot of these four ws into their system and they lived in a socially transparent world when we wrote this paper. So that's why we were able to talk a lot about these transfer cases challenges because those are some of the challenges we faced in the real world when we were trying to implement this in an enterprise setting that is multinational.

Andrey [01:13:24] I see. So when you presented this and sort of said, we should do this, you know how receptive our people today sort of get it right away or

Upol [01:13:34] initially there was a little bit of like hesitation, I think, because someone said, like, how is this explainability, right? Because there is this and there is a very powerful like AI developer like this is not explainability. And I think that's kind of like the idea of the paper kind of came to light. Our idea of explainability is so narrow that we have a hard time kind of even envisioning more than that. So what we actually did to kind of address those kind of concerns is, you know, as we see as you saw also on the paper in this empirical study that we had concrete like directly from the stakeholder information about how these additional context help them understand the system, right? And then if we go back to our initial definition of explainability rights, things that helped me understand the AI systems, right? And in this case, the AI systems are not algorithm. These are human AI assemblages. Right? So and they're socio technically situated.

Upol [01:14:36] So there you go. So initially, there was a lot of pushback, but what the proof is often in the pudding. So when we added social transparency, the engagement went from like two percent to ninety six percent. Right? That you can't ignore. Right. And so so those are some of the things that helped a lot of the stakeholders have more buy in and get a sense of, OK, now this is important. This might not look algorithmic, but it has everything to do with the algorithm, right?

Andrey [01:15:09] Yeah, I guess it harkens back to the title of a work, right? Expanding explainability, you know, as person said, how is this explainbaility while you've pointed out, then you sort of make the argument that this should be part of expandability, and by adding it, you get sort of a more holistic, full understanding. Is that kind of a fair characterization?

Upol [01:15:32] Yeah, yeah. And I think, you know, sometimes the simplicity is kind of elusive and deceptive. But, you know, we also have to understand that sometimes very powerful ideas and also very simple ideals. And I think within AI, we have to kind of go back to those roots at some point, like not everything that is complex is good. Neither is not everything that is simple is bad. You can have very good ideas that are very simple.

Andrey [01:15:57] Yeah, exactly. Simple ideas can be very powerful. And I guess one of the key insights here is social transparency as a concept and as something that needs to be part of expandability. So just to go back and situate within the XAI research field, you know, I don't know too much about the context of that field and what is going on there. So what do you think could be hopefully, I guess, the impact and what this could enable as far as future research?

Upol [01:16:30] First of all, I think it makes this very nebulous topic of socio organizational context tractable, right? Like for concrete things to go for, and that's a good starting point. It gives people to grasp on to that and build on it. And I think that's what we actually invite people to do right is now that we have at least started the conversation. That explainability is beyond algorithmic transparency and given the community one way of capturing the socio organizational context, I think now it starts to seed more ideas. And I think there is some fascinating paper that I've seen after that around and ideas actually that talk.

Upol [01:17:15] Using this notion of social transparency talked about end to end lifecycle perspectives within explainability, like who needs to know what, when and why, like sheep and ocher and Christine Wolfe and others have kind of written about it. So I didn't get it. It gives us bedrock for future work to kind of build on it, and I hope it does, and I'll work within explainability takes far beyond social transparency. There are other things that are

outside the box that also need to be included. And how do we encode that? I hope people use this kind of scenario, these design techniques, and it is also not shy away from the fact that if something as simple, right, as long as powerful, that's still a valid and good contribution.

Andrey [01:17:59] Yeah, I guess in a sense, that's how you want research to work, someone reads a paper and it's like, Wow, this is cool, but what if he did this or this thing doesn't work? You know, I have this idea. So that makes a lot of sense. And also to that notion of sort of the context and the field itself, we talked about on a bit of a push back, it got at the industry level within the research community. You know, when you submitted it, when you got reviews, when you presented it, what was the reception of your colleagues?

Upol [01:18:32] I think it was surprising to us. We always thought when we wrote the paper that people either hate it or they will love it. I don't think anyone who's going to be neutral to it because it was making a very provocative argument. It was making the argument that explainability is not transgressive. It is more than that. And it's not just saying that it's like this one way of doing it. So and clearly it was well-received, and the presentation at Chi went very well. And, you know, we were very lucky to receive this paper, honorable mention on it as well. So I think overall, it went better than we expected it, to be honest.

Andrey [01:19:13] Yeah, it's good to hear that given again, this was it looks to be quite the effort. CHI is pretty big, right?

Upol [01:19:22] Yeah, it is the premier HCI conference. So like, not like nervous for now because Neurips at a different scale, but like in terms of like the premiere venue. Right CHI, is that for HCI, what NEURIPS is like for ML.no, I guess that's a different way of looking at it.

Andrey [01:19:36] Well, so yeah, that's really cool. And we'll have a link to that paper again and a description and our Substack. So if you do want to get more into it, you can just click and read it. And that's just to touch on a bit what has happened since. In your research, you had actually a couple of weeks. So first up, you have the WHO and explainable AI how AI background shapes perceptions of AI explanations. How does that relate to your prior work and endless work? And sort of what what was what is it?

Upol [01:20:13] Absolutely. So I mean, this is directly related to a human centered, explainable way. I kind of work in the sense that not all humans are the same when it comes to interacting with the AI system. I don't think anyone will challenge that observation. Right. But then the question becomes, OK, who are these people? How do their different views or characteristics impact how they interpret explanations? So in this paper, it's just something that we looked at like a very critical dimension, which is any AI background. Like if you think about consumers of AI technology versus creators of AI technology, oftentimes consumers don't have the level of AI background that the creators have, right?

Upol [01:20:57] So given that this background is a consequential dimension, but also the fact that it might be absent in the users of systems that we build. How does that background actually impact the perceptions of these AI explanations, right? Because again, we're making the explanations also for the receiver explaining that and then they explain that so that this is the paper that is, I think, the first paper that kind of explores the

AI background as there's a dimension to to see like, well, how does that impact like we see humans, humans, but who are these humans? Well, let's look at two two groups of humans like people with and people without. So this paper kind of presents a study based largely actually on the Frogger work now way back when to kind of get at these questions.

Andrey [01:21:48] Yeah, it makes me think also, aside from like, you know, I develop or not add it all up, or even just like programmer who were and resources a person in sales. You might interact with the AI system differently, so it seems no good to take into account, for sure. Yeah. And then I think also you had this elevating explainability pitfalls beyond dark patterns and explainable AI, which sounds a little bit exciting. So, yeah, what's that about?

Upol [01:22:24] So this paper is actually related to the WHO in my paper, because one of the findings in WHO and say that we got was we're both groups. The group with AI and NONYE backgrounds had exhibited unwarranted faith in numerical based explanations that had no meaning behind them, so to speak. But even if people did not understand what the numbers meant, there was a level of over trust in them. So based on. That observation, what is interesting is like we were not trying to trick anyone, right? Like that's the importance of this finding that in the study, we were not trying to trick anyone. We just use the numerical explanations as a baseline. Our main instrument was the textual explanations, the actual rationale.

Upol [01:23:15] And while trying to examine that, we were like, Oh my God, why are people like so in love with these numbers, then that they don't understand? Because we have qualitative data where they tell us, I don't understand it now, but I can understand it later. And what is interesting is that people with AI background and those without. Have different results for over trusting the AI, right? So over trusting the numbers. Excuse me. Yeah. So we started asking the questions. All right. There are many times where harmful effects can happen, like over trust, even when best of intentions are there, like in our case. Right. A lot of harmful work and explainable AI is couched under this term called dark patterns, which are basically deceptive practices. It's easiest to explain it from the other side, like if you think about like, you know, in certain websites, they have all these like like transparent like ads. And when you're trying to click the play button like 10000 windows, open up, right? And you have to take them 10000 politicians to get it. So there's a dark side that kind of drives clicks by tricking the user, you know, not all harm patterns like harmful patterns are created equal.

Upol [01:24:31] So what happens when harmful effects emerge, when there is no bad intention behind it? Right, right? So to answer that question, we wrote another kind of conceptual paper, and we call these things explainable the pitfalls. Right? So these pitfalls are certain things that you might not intend for bad things to happen, but like a pitfall in a real piece of like in the real world, you might inadvertently fall into it. Right? Because, you know, it's not like pitfalls have there to like trap people. Sometimes the pitfalls emerge in nature, in jungles and other places by the construction site, and that you might inadvertently fall into it. So this paper is kind of trying to articulate what are explainability pitfalls? How do you address them? What are some of the strategies to mitigate them? So this is more of another kind of a conceptual paper situated with a case study, and it recently got into the human centered AI workshop at in Europe. So this year, so we are looking forward to sharing it with the community as well.

Andrey [01:25:32] Oh, it's exciting. Yeah, that's roughly in a moment, right? Yeah, yeah. Yeah, that's that's interesting. This concept of sheer is something you should avoid doing seems like a good idea, almost publishing negative results, which is which is fun. Well, we went for a lot of your work and then almost traced from the beginning to the present. But of course, it's also important again, to mention, as you have done before, that this was, you know, a lot of this was done with many collaborators and you built on a lot of prior research, obviously in many fields. This is true of any research job because you were present, maybe beyond your papers. What kind of is the situation when it comes to community working on XAI, Nick's family AI and also human centered XAI, you know, is is is your being human centered or socio technical? Is that becoming more popular or are more people so aware of it, that sort of thing?

Upol [01:26:43] No, you're absolutely right. I think, you know, I stand in the shoulder of giants, right? There's no two ways about it without the fantastic people I work with. None of this work becomes reality and the communities, and it's something that I care deeply about. So we have been very lucky in this context. And by 2020 one, we were able to host the first human centered explainable AI workshop. It was actually one of the largest attended workshops and trials during more than 100 people came over 14 countries. So we had a stellar group of papers. We had a keynote from Tim Miller, an expert panel discussions. So I think that community is still going on. And actually, we did just propose to host the second workshop at Chi. And I think after this, we want to take it beyond. We want to take it down. Europe's who want to take it to triple AI, to try to see how more can we intersect with more other communities around HCI, like other relevant social groups, right? The computer vision people and this people. So. These are some things that we deeply care about, and that is something that I would that I'm kind of like looking forward.

Andrey [01:28:07] Yeah, definitely so. And just to get it a bit more into that, you know, what's next for you both in terms of this community aspect of, you know, having various events to let people know about to see you and also in terms of, I guess, where your research is headed.

Upol [01:28:25] Yeah, I think for me, I as I share, if there's a project that I'm doing with radiation oncology, it's actually exploring social transparency in their world and this has been actually a value long term engagement. I've been working with them for more than two years now. I've also kind of been working with the Data and Society Institute on Algorithmic Justice Issues around the Global South. So you know what happens when we all talk a lot about algorithmic deployment right before deployment? Dataset creation? But what happens when algorithms get taken out and what happens, then what happens when they're no longer used?

Upol [01:29:04] So there is a project that I'm running that has explainability component, as well as algorithmic justice component around being creating the algorithmic trading of the DC exams, which are like basically international exams administered by Ofqual and UK governing boards. But these exams are actually administered in over one hundred and sixty countries. So you might recall that in August of twenty twenty, there were protests around an algorithm grading a lot of students know. While the reporting was great. It only focused on the U.K. we really don't know what happened in the other one hundred and sixty countries where these exams were administered. So, you know, beyond, you know, as I say, denies you kindly shared my bio, right? What happens to the people who are not on the table? And I think if you don't amplify people's voices, we're not at the table, they often end up on the menu. So I think coming for the circle like that, something that I'm

deeply curious about, so that's roughly like, you know what things are and if I have the privilege of giving a keynote at the World Usability Day actually tomorrow on November 11th, I have some invited talks lined up at the University of Buffalo on the 30th and then an expert panel discussion actually at the university's medical school, the Stanford Medical School, to the conference. So that's that's pretty much like like a ramp up to the end of the year.

Andrey [01:30:31] Cool, yeah. Sadly, will release as five guests pass through 11. But will these talks be recorded or public? Could be.

Upol [01:30:42] That's a very good point. Thank you so much for asking. So I am going to check it. I wonder what I would recommend if the listeners are there. If you check out my Twitter, if they are public, I will be sure to make sure that they are published and shared widely. So as of now, I'm not sure which of these would be public versus not. But if they are, I will publish them on my Twitter. So if people are interested and I think we can also add links to them after the podcast.

Andrey [01:31:13] Exactly. Yeah. So you can look down that description. We'll figure it out and we'll have links to this and all papers and everything. All right. So that's cool. And then as I like to wrap up, after all this intense discussion of research and ideas and studies, just, you know, a little bit about you and not your research. What do you do these days? Or, you know, in general, beyond research, what are your main hobbies? What are your main interests?

Upol [01:31:48] Yeah, I guess I'm, you know, I've been I love to cook. I think that is something that has been during the stay at home and pandemic mode has been a blessing. I absolutely love European football or soccer. All my team is not doing very well. Manchester United right now, but I tend to. That is my escape and I also play this game called Football Manager. I have not like fantasy football, but it's like kind of like that where it's a very data driven engine. And that's how it comes up to like this game engine that kind of predicts the future. I'm going to simulate games. That is my escape in terms of all the things in reality.

Upol [01:32:38] But I absolutely a big fan of old school hip hop. So I listen to a lot of music. I, whenever I get some time, I do mix beats on my own time for my own enjoyment. I don't think I have a song called Account or anything now, but those are my ways of keeping sane.

Upol [01:33:01] But most importantly, one of the most cherished things that I do is mentoring young researchers, especially who are underrepresented, especially who are from the global south. So I'm very proud of all the mentees that have taught me so much throughout the years, like ever since 2000. I think 11 12, I've had the privilege of mentoring around 100 hundred people from many different countries in Asia and Africa and kind of guiding them through high school and those. That is something that like gives me a lot of joy actually, like whenever I get free time. That's actually what I do. And during application season, it's usually gets tough because we have a lot of requests to review applications because, you know, sometimes as you can imagine, life like the application. The statement of purpose is often a black box, right? And you don't know what to write. So that is one thing that I get a lot of joy from.

Andrey [01:33:59] Yeah, that's that's fantastic. I think we all guys share what a good deal of mentorship as Diaz's adviser for a reason, you know, as an assigned mentor. So it's it does feel nice to give back, and I have always enjoyed being a teaching assistant and these various things are always pretty rewarding for me. Well, that was a really fun interview. It was great to see or hear about this human centered AI as a researcher who talks of robots refreshing to think about people for once. Thank you so much for being on the podcast.

Upol [01:34:41] My pleasure. Thank you, Andrey. I so appreciate the opportunity to talk to you and an animal in a way to you. Talk to the listeners. Thank you.

Andrey [01:34:50] Absolutely. And once again, this is The Gradient podcast. Check out our magazine website at The Gradient dot com. To you, Earl. And our newsletter and actually this podcast at The Gradient pub that Substack dot com, you can support us there by subscribing and also share all of this review on this Apple and all these kinds of things. So if you dig this stuff, we would appreciate your support. Thank you so much for listening and be sure to tune into our future episodes.