Professor Manuele Reani - Human-like AI: Comforting... or Manipulative?

Welcome to this episode of Exploring the Humanities, Women's Voices from the Hong Kong Polytechnic University. Our podcasts allow us to showcase the exciting and innovative work being done by our colleagues in the humanities. Intersecting with fields as varied as aviation, mental and physical health, virtual reality, GenAI, design, neuroscience...

I am Renia Lopez from the Faculty of Humanities. With us today, we are very lucky to have Professor Manuele Reani, Assistant Professor at the School of Management and Economics, working at the Chinese University of Hong Kong in Shenzhen.

- Thank you, Professor Lopez, for inviting me here and to this beautiful studio, and for giving me this opportunity to be interviewed by you.

You are very welcome. Thank you for being here. To start with, could you give us a brief overview of what your research is about?

- My background is in psychology and computer science. Basically, I'm interested in how the design of artificial intelligence affects human perception and human behavior. Specifically, I'm looking at how anthropomorphism – anthropomorphic design in artificial intelligence - affects how people trust AI and how they perceive risk.

Tell us a little bit more, what is anthropomorphism?

- Anthropomorphism, this word comes from ancient Greek. "Anthropo" means human, and "Morphism" means shape. So basically, it's a design style that makes AI look like a human or behave like humans. We use specific cues for that. For example, the image of the avatar could look like a man or it could look like a robot. Or for example, we can change the name. We can call the chatbot David, or we can call the chatbot 12345.
- We also manipulate the language, verbal communication and non-verbal communication. For example, turn-taking, or those dots that you see when the AI is thinking, those kinds of cues.

We'll come back to that in a minute, but first I want to ask you, how long have you been working on this?

- Well, I've been working in the area of HCI – Human Computer Interaction – for quite a while. But in the last three years, I've been more interested in AI design, because I noticed

that more and more services embed AI systems into their platforms. And those AIs look more and more like humans. They use this kind of anthropomorphism to create human-like AI. And I think this is a big risk if people don't understand what they're dealing with.

So, about three years, you mentioned that you've been studying AI design in these avatars. Can I just take you back? What about before then?

- Yeah, before I was still working on human-computer interaction, I was looking at how to visualize statistical information, and how different visualization styles affect human perception. So I was still interested in the human side of computer science, but more about data visualization, information visualization and statistical knowledge, which is captured by plots, diagrams, graphs, and so on.

I would agree that most of our audience will find the current work that you're doing probably more exciting, more interesting. So then, tell us, why that shift to AI design and human perception?

- Yeah, as I mentioned before, I noticed that this kind of design – the humanized AI, the anthropomorphic design – is used more and more in chatbots, in any kind of service. And I believe, actually, there is some proof of that, that this kind of design, in certain domains, like high-risk environments, for example, finance, healthcare, or even mental health, can actually be dangerous, because it can persuade people to do actions that they actually don't want to do. Or maybe they over-trust the output from LLMs – a larger language model – to the point that they don't even check whether the output is accurate, or it's just a hallucination.

Do you think the designers of these products are aware of these risks? Are they doing it on purpose?

- I'm not entirely sure whether they are aware of the risk. I think they have a clue about it, but the purpose of the designer is obviously to sell, to make money. So anthropomorphic design has been shown to increase engagement and adoption of these technologies. So they're more focused on the business side. And that's where I think regulators should intervene.

Definitely. So let's go into the details of how to make an avatar more human. You

mentioned before the verbal and non-verbal behaviors. So let's go into the verbal ones. What are the things that an avatar does that make it sound more human?

- Yes, that's a good question. So for the verbal communication, communication that resembles human empathy, for example. So pretending to understand the user's emotion. For example, a user might come with some problem, "I suffer from depression", or "I just broke up with my husband" or whatever. And the chatbot will say something like, "Oh, I'm sorry to hear that. It must be painful. I can feel what you are experiencing," or things like that, where the chatbot actually tried to use empathic words and sentences to create closeness to the user.

Do you think this is working now because we are not used to it? Because this is a novelty, basically, to have a chatbot talk to us like that. Do you think that with time we might get bored with it, or we might realize that this is the answer I'm getting all the time? For example, when you interact with ChatGPT, it will always tell you, "Oh, that's an excellent question. That's an excellent idea." I personally get bored with that. So do you think that might still happen?

- I think in a way, consciously, we might get used to it. I mean, we might dissociate from this empathic language. It's an effective style of communication. But implicitly, at an unconscious level, this might still have an effect. We don't know that. We won't know that for a certain time – time will tell, right? This technology is fairly new, and this design is also fairly new. In the first tier of ChatGPT's release, it was not talking that way. They increased the anthropomorphism recently, exactly because they found that it's more entertaining, more engaging, and more useful for business purposes.

Are there any studies that you know of where they have actually tracked whether any specific behaviors that are helping to engage their users and sell their product?

- Yeah, there are other studies talking about alignment, for example, effective alignment, how the AI recognizes and knows your feelings, your emotions, your understanding, how this can persuade you to do a certain action. Even dangerous actions like disclosing personal information, for example, are one of the areas where risk is very high.

[01:07:07:10 - 01:07:49:13]

And is this the area that you work on specifically?

- Also, this area I'm more interested in recently, I'm more interested in financial applications and healthcare applications, but data protection is also another area that is very important. And in marketing, we know that this is kind of a tricky business because with the cookies before, now with the AI, the marketers always try to collect user information, private information, for marketing purposes.

Tell us about the work that you do with finance and how these avatars are used in the finance industry.

- We know that now more and more avatars, more and more AI are used in advising. So for example, financial advisors before real people, now more and more we see these advisors being replaced by AI. And these advisors are basically chatbots, smart chatbots that give you advice on how to invest your savings. And I study how the manipulation of the interface, the anthropomorphism of the interface can affect risk perception about an investment. And this is a very dangerous area because as we know in the past from past financial crises, we find that people were executing transactions based on hype. So they were buying stocks or they were purchasing financial products based on optimism, hype over the market, or over some companies. And then this turned out to be a disaster in certain cases. So I'm looking at how the manipulation of the design of AI can persuade you to do things that you would not normally do.

And you've actually done these studies with real people looking at real financial products?

- Yeah, the way we did it is with real people. We ask people to go to a platform, an online financial advisory platform, and ask about how they should invest their money, their savings. And on the other hand, we have a chatbot talking to them and giving advice and also explaining why that advice is given, like transparency, transparent AI. And we manipulate the style that communicates more empathetically, more warmly, closer to the user.
- Also, the avatar, we can use like a picture of a computer or a picture of a person. We also manipulate the name. So we use David, for example, for the anthropomorphic one and Bot 12345 for the non-anthropomorphic one. And we check how this manipulation affects the perceived risk, but also their trust. Because trust is very important here.

And tell us, what did you find?

- Yeah, I found that this manipulation of human likeness has an effect, in fact, on how people trust the AI. It increased trust and consequently, it decreased risk perception, which is kind of a dangerous thing, but only for users who are laypeople. Experts actually had kind of the opposite effect. I think for experts, people who are really knowledgeable about finance, they perceive this kind of empathetic language as deceiving.

Going back to what I was saying before about the ChatGPT telling me everything I do is excellent when I know it's not.

- Exactly. And this is found in experts; probably there's some explanation in behavioral economics. Humans are not rational. And when they see emotional content into a chatbot, they might perceive some irrationality there, right? Or some sort of deceiving kind of language.

Have you found any difference by gender, by age?

- That's another interesting point. Yes, we found not so much about age, but with gender, we did indeed find some sort of differences. We found that male users tend to be more deceived by the AI, tend to anthropomorphize AI more, especially when the cues are not so obvious. Female users tend to be more resistant. They know they are interacting with the machine. Obviously, when you anthropomorphize the chatbot a lot, then the difference is basically non-existent.
- We don't know why this is the case. There might be some developmental psychology theory that can explain this, but it's all speculation. We actually don't know why this is happening.

And was this moderated in any way by the level of knowledge of the users?

- The expertise, of course, has an effect on this, but it's non-linear. So it's quite complex. We cannot really give a recipe for this interaction because it is context-dependent. There are a lot of factors that can interfere, and it's non-linear. So it's difficult to pinpoint exactly what's going on.

And out of all the variables that you were playing with, was there one that made a bigger

difference than the others? So you've mentioned the name, you've mentioned the picture...

- Yeah, so out of the cues – the anthropomorphic cues - I think images, visual cues, are more effective. But even that depends on the interaction. Even the communicative cues are very effective. The least effective is what are called identity cues. For example, the name. The name is not really that important according to our results.

And you're talking about images, the photographs, the images. I'm assuming that these were images of real people, or they look like real people. What if you were to include other behaviors, like those related to movement? Do you think that might make a difference as well?

- We didn't test that, but definitely future research will test it. Because I think the movement of the avatar is definitely very important. It gives even more of a sense of human likeness, in my opinion. Maybe you can comment on this, given that you are an expert in this field.

I know, but I'm not being interviewed here.

- All right, all right. I think definitely if this is something that you're interested in, you can definitely explore this area. Because the more cues you use, probably the more you give the sense of humanity to the AI.

I'm also thinking about gaze, for example. Even smiling.

- Or blinking.

Or blinking, absolutely. Yes. Okay, so definitely an area for research there. So in some of your papers, you talk about the fundamental over-attribution error. Tell us, what exactly does that mean, and how does it tie in with everything that we've just been talking about?

- Yes, so the fundamental over-attribution error, or FOE, is a new bias. Fairly new bias. It was defined in a book published five years ago, but then it was not really explored by research. Basically, it means that the user over attributes agency, intention, emotion, to an AI, to the point that they believe it. So there is a difference here between anthropomorphism and FOE. Anthropomorphism is a human tendency.
- There are theories that explain that we evolved to develop this tendency. So basically, they

say it's evolutionary fitting. It serves a purpose, which is to enable us to interact with the environment. So we tend to talk in a human way to objects, or even to animals. Think about people who have pets.

Yes, yes.

- And they talk to them, right? Or they dress them like babies sometimes, right? So we tend to do that, and that is useful for us to be able to interact effectively with the environment. But FOE is not useful. Actually, it's not a tendency. It's a bias. It's a bias that doesn't help us achieve any goal. Actually, it can lead us to unwanted behavior or detrimental behavior.

And this is something that, from what you're saying, some industries are exploiting for their gain.

- I believe so. I believe they exploit this FOE for persuasion. It could be disclosing information. It could be buying products. It can be anything.

Do you see a positive side to it?

- I don't think FOE is positive. I think the definition of it is exactly that. It's not positive because it's a bias. It's a misperception of reality. And whenever you don't see reality for what it is, it's always negative.

I'm thinking in the health care sector, for example, is this something that could be exploited for a positive purpose?

- Anthropomorphism, yes, definitely. Using anthropomorphism in communication for mental health can be very beneficial. But FOE, to the point that the user believed that the AI has intention, positive intention, that the AI cares about the user, that's not really beneficial. Think about depressed patients who go on ChatGPT and ask for advice, how to deal with the relationship, for example. And then they follow those advice without double-checking, whether the output is a hallucination or is bad advice. This is a kind of over-trust, right? And this can be very, very bad, bringing very negative consequences for depressed patients, for example.

Are you also looking at this type of behavior in the healthcare?

- Yes, yes. Right now, currently, I'm developing an agent, and it's an AI coach, AI counselor –

which uses this sort of language and different design styles, including anthropomorphism, to advise. I wouldn't say patients because it's more about stress for the exam, so we don't have severe depression. I'm not entering that domain, which is very tricky, but I'm using students who are very stressed over exams, or even managers who are willing to perform in the workplace and have a lot of pressure. So it's more like coaching, not really psychotherapy. But yes, we are working on this and manipulating different design styles to see the effect on the perception of the user.

Any results so far?

- No, we didn't collect data yet. We're just working on the design of the agent.

Because we were actually talking about that with schools. As you know, Hong Kong has a very high rate of suicides, and one of the ideas is that this interaction with a friendly AI agent might help to pick up or identify students who might be at risk before anything happens, basically.

- That's possible. The issue is that should the AI disclose this information to some government, because this is a private communication, right?

Yes.

- What is the extent of privacy here? Because you can use this technology to prevent suicide, but at the same time, you violate some privacy concerns. I wouldn't like the AI to tell the government about my conversation with it, to be honest.

So in that case, how could one control what the conversation is and how the agent is actually talking to the patient or to the person?

- I think for some clinically proven depressed patient – diagnosed depressed patient – we should have a different chatbot, we should use something more controlled in a controlled environment, they should interact with an AI which has been designed for this type of person, not with the general LLM that everybody else uses. And then in that case, there could be some informed consent where they accept that if there is some dangerous communication, this communication might be communicated to the doctor or the hospital or whoever is responsible for it. Then it's possible. And I'm not sure whether the patient would prefer to use this kind of AI or the general AI models. That's another story,

but definitely we can nudge, we can persuade this to happen.

And why would a patient prefer to talk to an AI agent rather than a person?

- Yeah, there should be some advantages, right? So I think using some framework, for example, a CBT framework or positive psychology framework – a science-based framework for psychotherapy – would be an advantage. So if you market these AI psychologists as an evidence-based AI coaches or AI counselors, then it would be more persuasive for the user to choose something that has been evidence-based design rather than a general-purpose model. So it needs to be marketed that way.

We're looking at the moment at interactions online and face-to-face, and we're actually finding quite a number of differences between the two, as we probably expect. But this goes beyond visibility because in our interactions, people can see each other fully. And we are thinking that there is something else going on in the face-to-face interaction. Are we going to be losing out on that if we move all of our contacts to online, basically, to AI?

- Yes, I think so. I think we are losing some communication cues that are implicit and difficult to replicate with technology. At the same time, being a psychologist, a therapist is very hard because as human beings, we come with our own baggage, right? And when we interact with the patients, we reflect, even though we don't want to do that, we tend to reflect those traumas that we have onto the patient. So we react.
- So the advantage of AI is that AI does not react – doesn't have trauma. So an AI therapist is neutral 100%. Or at least, depending on the design of the algorithm, but it should be in theory neutral – definitely does not react based on trauma. So that's one advantage. But of course, we are losing out on the communication cues that you have in real-person-to-person communication. That's definitely one of the problems.
- But like your research is trying to do, right? You are using movement and blinking and all this, we can get close to it. We can get really close to it with time. I think right now we are not there, but research is going in that direction.

Picking up on something you've just said and linking it to the work that you do with the finance industry. Now, can we not look at the positive side of using AI as an expert finance advisor because it doesn't have that emotional element to it because it doesn't react?

- Definitely. In sectors such as finance, having AI advisors might be actually be better because there is no irrational behavior. The advisor is looking at the data. It does not react. It doesn't have an emotional reaction or anything like that. And you can throw anything you want at the advisor. It will not disqualify or make you feel bad and create some emotional reaction in yourself. So definitely using AI for finance, very technical. I think that even law - the legal services – might be beneficial. But when we talk about healthcare, we need social presence. And robots and AI, they're not made for that. It's very hard to bring social presence, human presence into a hospital, for example. For mental health, it's even worse because we're dealing with psychology here, not just human contact, but also you have the empathy, the understanding that a human being can express with the whole body, right? That's something that you cannot really replace with AI. As I said, there are some advantages of AI, which is the one I just described, the neural reactivity based on trauma.

And one last thought is that we have been communicating face-to-face for thousands and thousands of years. And now all of a sudden, in the last couple of decades, we have completely changed the way that we communicate. Comment on that!

- In the last 20 years, after the smartphone came about, you can see that there are a lot of people with mental problems. Mental health deteriorated in the last 20 years, for sure. I think this is a trend that is going to reach a plateau. I think more and more people would like to use AI and to interact with technology rather than human beings, exactly because they don't react. They don't have this emotional reaction based on trauma. So sometimes it feels good, right? I think you should try for yourself. You chat with ChatGPT or Gemini or whatever, and about the problem, right? Maybe you chat with your sister before, you feel better chatting with the AI sometimes because your sister might judge you, might use morality over you, might use some language that is a reactive emotion, right? And then you chat with ChatGPT or Gemini. You don't have that. So for certain communication, for certain topics, AI feels better. But we risk, as we said, to deteriorate mental health more and more. I think we will reach a plateau very soon.

And can I also ask you, what about AI in education? What are your thoughts about that?

- Yeah, that's a controversial topic. There are a couple of papers which came out in a very influential journal, which is called “AI in Society”. I think you know it. And they talk about

bullshit, right? Bullshit and hallucination. They compare hallucination to bullshit. They say that LLM hallucinates quite a lot and this is a form of bullshit. So it should not be used in education.

Also, one problem with education and AI in education is that we don't want our young students, especially people who are in elementary school, middle school, to start using phones in class. Because right now they're not allowed to do that. But if we allow the use of AI, how can they use AI? They need to bring the phone into the classroom. They need to use the phone. We want to discourage that. That's another problem. So it's a very controversial topic, but there are some advantages of AI in education as well.

Are there any last thoughts that you would like to share with our audience?

- Yeah. So I think all this research and all this discussion is not made for scaring people. We don't want to scare people and to discourage the use of AI. I think everybody should use AI for everything, for every purpose. It is a very useful tool that has changed in society and people's lives quite a lot, especially after the generative AI revolution. But I think governments and regulators should step in and consider these problems over trust, FOE, risk perception, design of human likeness into AI and put some limits on what those developers, those companies can and cannot do. And also educate people about these problems. So warn people of what AI is and what it is not in a way that even though they know they're interacting with an artificial agent, this is being reminded over time. That's definitely beneficial. Yeah. So don't be scared, use AI, but be mindful of what it is.

Thank you very much, Professor Reani, for your thoughts and also for your advice on how to be careful when using AI.

- Thank you, Professor Lopez, for inviting me here and for this lovely interview.

Thank you for joining us on exploring the humanities from the Hong Kong Polytechnic University. For more episodes or information, visit our website or follow us on Spotify. Stay tuned for discussions with leading voices from the Faculty of Humanities and beyond.