There is method to the madness

Transcript of the Human-Robot Interaction Podcast Episode 16. Published on 18 December 2020by Christoph Bartneck at https://www.human-robot-interaction.org/2020/12/18/there-is-method-to-the-madness/

[00:00:00] **Christoph:** The success of human robot interaction depends on utilizing our understanding of the interaction between humans and robots in the development of new technology. These new technologies then need to be rigorously tested to prove their benefit. In other words, it is time for an HRI study. In this episode, we will discuss some of the major decisions you need to make when designing a study. Choose wisely for while the true study will bring you knowledge, the falls we'll take it from you.

[00:00:34] This is the Human Robot Interaction podcast. I'm your host Christoph Bartneck. Tony Belpaeme recently published a book chapter entitled "Advice to New Human-Robot Interaction Researchers", in which he discusses many of the choices that you need to make for an HRI study. Tony, welcome to the show.

[00:01:06] **Tony:** Thank you for having me.

[00:01:08] **Christoph:** Come on. You can do that with a little bit more enthusiasm.

[00:01:10] **Tony:** Ha, Ha! Sorry. I'm just reading the script now. Like welcome. Say it again. Welcome to the show and I'll

[00:01:18] **Christoph:** Tony. Welcome to the show.

[00:01:21] **Tony:** Thanks for having me. I'm a great fan. Really! It's an honor to be here.

[00:01:26] **Christoph:** When I invited you to talk about this paper, you told me that you initially wrote this as and I caught you here "as a throw away chapter". Why do you think so little about your own work?

[00:01:36] **Tony:** Well it's not that I think little about my work, but I was asked to write this chapter to go into a book and, academic books are not sold in their millions. So I felt that my pearls of wisdom might be lost in a book. And, if the book was not read, not bought by anyone, now it turns out that book is actually rather brilliant. And so lots of really interesting contributions by my colleagues who I valued very highly. And so it was great to know that you at least read it. So that's, Hey, that's one reader, so I'm very happy.

[00:02:04] **Christoph:** What motivated you to write this chapter?

[00:02:08] **Tony:** I just want to write about the struggle of doing experimental work in human-robot interaction. And I wanted to share some of my personal insights. So together with the team, we've been doing studies in human-robot interaction for over 10 years. And as I'm an engineer and a computer scientist, it means that I'm always I'm a bit challenged when it comes to experimental work. Yeah. A psychologist they've been trained in the scientific method, but geeky people such as me are not. And so I've made pretty much every beginner's mistake that can make.

[00:02:42] **Christoph:** After I read your paper, several additional high level decisions came to my mind. Normally I would give the guests the honor of bringing up the first issue. But I believe that at least two questions need to be answered before addressing the issues you raised in your paper. Would you allow me to bring these up first?

[00:03:00] **Tony:** Oh, yes, please. I would love to hear your two questions.

[00:03:03] **Christoph:** You have chosen wisely. The first question that I usually bring up is whether you are trying to solve a problem, like an engineer. Or whether you are trying to answer a question like a psychologist. The first is trying to improve the world while the second is trying to understand it. These two approaches are not mutually exclusive since understanding a problem is a prerequisite to solving it. And nothing is as practical as a good theory. Still far too often, are we presented with solutions that are looking for problems and answers that seek questions. Furthermore, there are many problems to which technology is not the solution. Tony, have you ever reviewed papers of such a nature?

[00:03:46] **Tony:** So I've seen many papers where the authors throw technology at the problem. But somehow a low-tech solution would have been better and I've been guilty of that myself. So for example, some recent work where in the COVID-19 pandemic, we've seen robots pop up that measure your temperature.

[00:04:05] Basically, that's a thermal camera mounted on the robot. The robot is just a mount for a thermometer. Why is that robot there? Does that problem really need a robot? Yeah. All too often we have engineers and scientists that get carried away and they want to use up their favorite tech in my case robots.

[00:04:22] And we want to use that technology to solve problems that perhaps don't really need a robot solution, to a man with a hammer everything looks like a nail. And it's not just us as robot builders we sometimes think that robots might be the solution to a whole raft of problems, but also people really think that we have solutions to their problems and that the very latest technology is exactly what they need

[00:04:46] Nowadays, for example, every problem seems to be solvable using artificial intelligence and robots. The snake oil salesmen are already there to exploid that belief. I had people knocking on my door believing that robots could cure Alzheimers and wanting me to build those robots, you know, While robots were able to do many wonderful things, curing Alzheimer's or cancer is really not one of them.

[00:05:07] **Christoph:** Engineers should define the success criteria for the solution before they created it. This criteria should be specific and measurable. Moreover, it should also consider other existing solutions, including those of human performance. Is this solution better than a human doing the same task? Often I see studies that compare two variations of a certain technology. And usually the one of the authors performed better. Comparisons to other robots and other research solutions are rare. Barely ever is the performance of a robot compared to that of humans. Do you know why?

[00:05:43] **Tony:** I've got an idea, say, yeah, you're right. If engineers wanted to demonstrate how good their solution is, they will often contrast it to something else. And that is good practice, of course. To compare what you have to do something else out there. Sometimes there are benchmarks out there or people will report their results and you can compare it to that established solution.

[00:06:02] Sometimes there's a treatment as usual. This might not be the best solution out there or the best possible algorithm or the best possible robot, but it might be a common standard in your discipline

[00:06:13] You could compare, I don't know, a computer vision algorithm to one that can ship with your robot to do something that is already popular, that, the entire field uses.

[00:06:23] But sometimes, ever so often people will compare their work to a straw man solution, just an earlier version, for example, of their robot or their software. This is not always bad. Yeah. Sometimes you really want to compare it to something that was before or you want to compare it to something that is not perfect.

[00:06:40] So you want to really. Break your end solution and then compare it against that. For example, the neural networks or ablation studies, where you break things in a neural network are used quite often to see what damage, what that damage does to your performance. You deactivate parts of neural network and see what that yeah, what influence that has on the network's performance. A robot has many moving parts and in a very literal and figurative sense too. And breaking something on the robot is sometimes a good way of seeing if the performance goes down. But as you said, we tend to report results compared to a worse solution, and we never compare it really to the performance that people have.

[00:07:21] Robots are meant to partially take over or compliment people. And so of course, we need to ask that question yeah. People are the golden standard. So how close does the robot solution come to what a human could do? And, and a Turing test, for example, can quantify how close a system comes to human performance.

[00:07:39] Turing test isn't only for chatbots, you can use it on different skills. For example, how human-like machine sounds or how human-like machine moves that can be tested using such a comparative method. But the holy grail still is to see if robots have the same application outcomes as people.

[00:07:58] And for example, if you build a robot tutor for teaching maths you want to ask, is it as good as a human tutor? Or if you build a robot as a friend, you want to know. Is that robot as good a friend as a human friend would be. And we don't often make that comparison. I think for two reasons, I think one is that comparing a robot to human signals that it is ready to replace human skills.

[00:08:23] And we do we worry about that. So there's this promethium danger to this. So building something that competes with human skill, especially if that thing, the robot resembles a human that is always met with caution restraint, where we don't really like doing that. And the other reason is that we just know really that the human is going to be better. We know that, there is no competition. The human will pretty much on every front outperform the robot.

[00:08:51] **Christoph:** When it comes to social interaction, humans are very difficult to outperform. I once supervised a student that openly admitted that she did not want to compare her system to the golden industry standard or that of a human performance. Since this would set her up for failure. She eventually graduated, although not under my supervision. Tony, what do you advise your students when it comes to the definition of success criteria?

[00:09:19] **Tony:** Depends a lot on what the topic of study is. For example, at the moment in the lab, we're looking at using machine learning to drive the nonverbal behavior of a robot. So we have a data set of people talking, and we feed their words and gestures into a sequence to sequence network, and that then generates nonverbal behavior for the robot.

[00:09:37] And one way to test is to test how good the network is. Is to use it to control a robot. Yeah. So the robot speaks a sequence that it has never encountered before, and it generates the matching hand and body gestures. And then you can ask people what they think of this. And usually you'll get some idea. People can rate how natural it looks, how human-like it looks, how fluent it is.

[00:09:59] But I think the best way forward is to do comparative study. And so to take, for example, gestures that you'd record it from people and replay these on the robot. And that is going to be your gold standard. You ask people to rate those, and that is the performance that you're trying to reach an alternative way is to really compare this.

[00:10:17] So show one robot that is replaying human gestures. Show another robot that is using computer generated gestures, and then asking which one is better or which one is preferred. And this is again, this like mini Turing test.

You can even go beyond that. It's not really Turing test. You can go beyond just recognizing which one is human and which one is machine generated.

[00:10:37] You can even ask which one do you prefer over the other and who knows, perhaps you'd there can outperform what humans do. But I think it all falls a stands with the quality of your human gold standard. If I build a robot tutor and I compare it to the worst teacher on the planet. What, if I dig up a real life, Dolores Umbridge, then of course your robot will come out on top so or if I asked if a robot voice is more pleasant than an irritating nasal voice, then that isn't a fair comparison. So you need to be very careful about what you compare your robot's performance to.

[00:11:20] **Christoph:** Okay. I have one more issue that I'd like to share before we go through your paper. Is it okay, Tony? Are you seated comfortably? Are you ready to give me a kiss?

[00:11:31] **Tony:** I'm ready for a kiss any time. Too bad we're just half a planet removed from each other.

[00:11:35] **Christoph:** We might be almost exactly on the opposite side of the world, so it's probably the furthest way a kiss could travel. I appreciate that. But what I've of course meant is. Keep it simple, stupid. Most researchers tend to make their first study overly complicated. They usually regret it when it comes to the analysis.

[00:11:53] And I'm no exception. Almost every additional experiment manipulation requires an increase in the number of participants. And particularly when you were investigating a weak phenomenon. Furthermore, you should never measure anything that you do not know how to analyze or to interpret. You need to be able to explain the relationships between all the variables in your experiment. Tony, have you ever included a measurement in your study that you later did not report on in your paper?

[00:12:22] **Tony:** To be fair. It has happened, so. So sometimes we just throw a measurement at a study and look what sticks. And so for example, we once assumed that the personality for people would correlate with how they responded to a robot. And so we had everyone take a personality test, using the big five personality test.

[00:12:38] And in the end it turned out that people's personalities really didn't predict anything about how they respond to the robot. So given the page limit we didn't report that in the publication. But it's a bit like this. I think that running studies is expensive. Not necessarily in money, but certainly in time, goodwill, human effort involved.

[00:12:58] And so it's understandable that you want to capture as much data as you want, while you're running the study. So even though you might not need the data to answer the research question at hand, you just grab whatever you can. Set up the cameras, roll video tape, whatever, just collect all that data.

[00:13:13] And perhaps it's not best practice, but I think it's just cautious to grab as much data as you can given that the ethics committee allows you to capture all that data. The only way to avoid that is to have a very clear idea of how to run your studies. I had to put focus in your study.

[00:13:34] **Christoph:** That brings up the question. How do you focus your studies?

[00:13:37] **Tony:** So I think the best ones are the ones where you start with a very clear research question and someone once gave me the advice that, before you start your research, start with the title of the paper you're going to publish on the research and. While I don't want to commit to a conclusion before running a study, it really helps to very succinctly say what the focus of your study is going to be about.

[00:13:59] And from that in a number of very clear questions and hypothesis, just follow and the rest how you implement the study just follows from that. So I think our most focused studies are really started with a single statement of what it is that we are going to study.

[00:14:28] **Christoph:** Okay. That is enough from my side. Let's talk about your paper. You raised the issue of either running a study in the lab or in the wild. What are the advantages of running them in the lab?

[00:14:41] **Tony:** When you were in the lab, you've got control over a lot of aspects, while you cannot control the individual differences in people, but you can make sure that they all get the same experience when interacting with the robot, you've got control over the room. You can avoid any distractions. You've got control over. I don't know how much coffee they had or how warm the room is. And so you can run a really tight ship in the lab. And so the lab is ideal to measure very specific things or very kind of small effects. So if something very minute that you're trying to pick up then the lab is ideal for that.

[00:15:14] The lab is convenient. We need to be honest about that. A lot of lab studies are just run because, Hey, you don't need to leave the office and to run a study and you don't need to jump through all the hoops involved in running a study in the real world.

[00:15:27] **Christoph:** The goal of HRI is not to enable robots to act in the lab, but amongst people in everyday situations. How can you generalize knowledge that was produced in the lab to situations in the real world?

[00:15:41] **Tony:** Studies in the real world are an entirely different beasts to lab studies and results from the lab are almost never replicated in the real world. They can be replicated, but people just don't do it. For example, if you want to know what the impact is, of eye blinking on the interaction, something, then that might be such a small signal that you can only pick it up in a lab study. And if you go to the real world, all the noise of the real world just washes out that results. So if you're trying to pick up a small signal, then the lab study is right thing to do, and sometimes you don't even need to go to the real world.

[00:16:14] So if you want to know if people like a female or a male sounding robot voice for your particular robot design, then you don't need to go into a shopping mall to ask people. You can just, I don't know, run a small study in the lab. And usually that'll give you results that are pretty solid. And won't change in the real world, because the real world is a mess. So you want to avoid running simple studies in the real world. There's a lot of external influences over which you have very little control and all these will influence your results.

[00:16:46] **Christoph:** How do you control all the random and unforeseen factors that exist in the real world?

[00:16:52] **Tony:** What happens is that all this noise and all this randomness adds variability to your results. And the only way to really deal with variability is to collect more data. So if you have a lot of noise, taking multiple measurements is going to help you pick up that signal in the noise.

[00:17:08] That really seems to suggest that I'm promoting quantitative methods, so where you measure data, but that's not the only thing you can do you so next to quantitative studies, you can also do a qualitative study in which you just observe how people interact with the robot, or you interview people about their experiences and their opinions.

[00:17:27] And somehow qualitative studies are considered to be the ugly duckling of empirical science. But I think in human robot interaction, they don't deserve enough attention. And certainly in experiments in the wild qualitative studies have real value.

[00:17:44] **Christoph:** Let's take the example of speech recognition. While it would work in a silent lab, it is likely to become unreliable in a noisy primary school. How do you enable robots to work in these demanding situations?

[00:17:59] **Tony:** You are absolutely right. So stuff that works in the lab suddenly goes tits up when you should go into the real world. And speech recognition is always one of the first things to go, to deal with that we can try and script the interaction very narrowly. So we can lead the conversation down a garden path of multiple choice questions, or we could use alternative input methods.

[00:18:20] For example, we use a touch screen to record the input to participants and quite often study subjects don't even know that the robot doesn't understand speech. For example, when we're working with children, we use combination of a tablet computer, and a robot, and the children will talk to the robot and they will believe that the robot understands them.

[00:18:38] But really the robot is only responding to what their fingers tap on the tablet. Sometimes technologies really fall short. If one of the children then suddenly asks an open question to the robot" what's your favorite color", then the robot can't answer. And then we have a tele operator, a wizard on standby to take over and quickly type in an answer on a keyboard that has then spoken through the robot.

[00:19:03] **Christoph:** But if you control the robot remotely, are you not deceiving participants into believing that robots have the abilities that they currently do not. Does such an approach, not systematically mislead society into a set of unrealistic expectations?

[00:19:20] **Tony:** Absolutely to some extent there is a degree of deception. Now it's a necessary evil to get at the data that we need in academic research usually that deception is revealed at the end of a session. So we explained to the participants that some or all of what the robot did was actually done by a human just pressing buttons on the other side of a wall or the other side of a panel, but outside academia, we're not always that honest.

[00:19:45] So especially when it comes to television, we don't reveal that the robot is remotely operated and that indeed sets wrong expectations. I had this personal experience, which I really didn't enjoy. So we recorded a TV program called "The robot will see you now" where we had a robot therapist interview people.

[00:20:06] And we wanted to see, how much people would reveal to the robot. It was great TV. It really was. Now we first tried to build this robot. We actually tried to build the AI for that robot and we didn't succeed. And then we moved to a wizard of Oz approach where people would try, someone was typing the answers that the robot was speaking on a keyboard, but that then was too slow for TV because, you want a quick television, you know, you want it to be entertaining. So we decided to hire the services of a Canadian voice actor. And she sounded right, like a robot and she could do a really convincing robot voice. And she did the Wizarding. Now we never told the people who stepped in that this robot wasn't real.

[00:20:49] And in the TV program, we never revealed that this robot was teleoperated, which I think was a missed opportunity. But worst thing was that after the TV show, I was contacted as I was the scientist on the TV show. I was contacted by people who asked me where they could get my wonderful robot and that they had, for example, I don't know, a suicidal son or daughter who would benefit from speaking to my robot.

[00:21:18] They all have to disappoint these people saying Hey, I'm terribly sorry, but the robot you saw in the TV program isn't real, it doesn't exist. Yeah. It's a mock-up something that we might have in the future. And it was great to see how people responded to the robot, but it isn't there. It was entirely tele operated and it broke my heart to have to do that.

[00:21:39] **Christoph:** That's indeed a sad story, but is it actually possible to build a robot to act fully autonomously in HRI studies? Or is that a bridge too far?

[00:21:54] **Tony:** I think it's a bridge too far at the moment. Then if you think about it, social interaction really involves every part of your brain, your language area, your memory, your motor areas, emotion, memory, everything is in synch. It is up and running in your brain when you're having a social interaction. And it means that we need to build an artificial equivalent of all these parts on the robot. And we are nowhere near that.

[00:22:31] **Christoph:** One of the major challenges for robots is to make sense of the world. This becomes much easier when they occupy a virtual world. Controlling an agent on the screen is much easier than controlling a robot in an unpredictable environment. Should we not first try out screen-based agents?

[00:22:50] **Tony:** Okay. You're right. Virtual agents they don't have batteries and that they don't run out or they don't have to deal with the reality of physics and gravity. And you could go beyond that. You could even build a virtual environment where physics is simulated. The way you have ultimate control over everything that happens. But the I don't know, I'm not entirely sure of that virtual agents are the real thing.

[00:23:12] **Christoph:** Here comes another, but isn't the final goal to operate robots in the real world. How can a screen based character help us in the development of real robots?

[00:23:21] **Tony:** I think to some extent screen-based characters could be good because development is cheap. Robots are, fickle and difficult and expensive. And so if you have a screen based character, you can develop faster than on a robot. And sometimes you can do experiments that are difficult to do with the real robot.

[00:23:38] Say, for example, if you're using crowdsourcing, so where you collect your data from people on the web, it's impossible to do that with a real robot. You can't send a real robot to, 500 people's homes. And so instead they have an interaction with a virtual robot, and I think this is a good second best.

[00:23:54] **Christoph:** I often wonder what the advantage of a physical embodiment is for a robot. We already have so many voice-based agents such as Alexa or Google assistant that worked just fine. If the main function of a robot is to be social, would such smart speakers not be sufficient?

[00:24:12] **Tony:** I don't think so there's something special about a social robot. There's something that makes our brain set up and pay attention and it's their appearance, their emotion. And it's just so much more powerful than a cylinder. Amazon Alexa from which a voice emanates or these robots can exhibit basic human social skills.

[00:24:30] They can respond to your presence. They make eye contact, they can point at things. And by doing that, they inhabit your social and physical world. And in a way that these voice-based assistants do not.

[00:24:41] **Christoph:** Not only robots can become virtual. Also participants can almost become virtual. As a matter of fact, Amazon is offering an application programming interface for their human workforce. You can integrate humans performing tasks into your computer program. This is often done to moderate forums. Humans' ingenuity for insults is no match for automatic filters. How do you use such crowdsourced workers in your experiments?

[00:25:11] **Tony:** We've never used it workers to handle some of the artificial intelligence of our robots. Although it does happen. Remizov at some of the pizza delivery robots in the States are actually remotely operated from Columbia, but the way we use crowd workers is to collect data from them. In studies, we use platforms like Amazon Mechanical Turk or Prolific to ask people questions about robots.

[00:25:37] Yeah. We might show things. We might show them little videos or pictures of robots, and we ask them to rate things. We asked them to, I don't know annotate data for us and so. In that sense, they are an online substitute for participants in the lab.

[00:25:51] **Christoph:** And what is the advantage of crowdsourcing experiments in this way?

[00:25:56] **Tony:** One of the greatest advantages for us is that you can get your results in a matter of hours. It's really easy to set up. Prices are ridiculously cheap, for under a US$, you can get about 20 minutes of work out of a worker on Amazon, which is so cheap, perhaps even too cheap. You can also get a much more varied group of participants from around the world. So if you want to know how Indian people perceive robots as compared to Japanese people and you're based in Canada. Easy. Yeah. Set it up online and you've got your data.

[00:26:24] **Christoph:** Needless to say that such an experiment could not use a real robot. Would such experiments there for not be heavily influenced by the preconceptions that the participants have of robots? The only prior experience with the robot would have been through watching them on TV.

[00:26:42] **Tony:** We have no control over which people join us in these crowdsource experiments, or very little control, really. And we don't know what their preconceptions are, so we don't even know what, when they're doing the work for us for all we know they might be, I don't know, on the tube somehow, or they might be sitting in the bar.

[00:26:58] I don't know. We have no control of that, but mind you, so everyone has preconceptions, not only the people who take part in your crowdsourcing studies, but also the people who walk into your lab to take part in the study. They have a particular idea of what robots can and cannot do to have a particular liking for robots, their perception of robots and their experience with the robot you're showing them is coloured through what they've picked up in the media.

[00:27:23] **Christoph:** Which brings us to the question of what participants to recruit. Most of the time, researcher opt for the convenient way, meaning students on the campus. Why is this a problem?

[00:27:36] **Tony:** Taking such a convenience sample. Isn't a problem. If what you measure is not really sensitive to age or socioeconomic status, but I think very few studies really are not sensitive to that. So in essence, in our lab studies right now, we're measuring what a 20 something highly educated, predominantly white person thinks of a robot interaction. And that isn't a good thing. Of course, there's no diversity at all in that. So we need to go out of our way to recruit a more diverse set of participants to our studies.

[00:28:14] Easier said than done of course. You can so beyond your campus doing experiments for credits like many psychology students have to do. Then you get a super narrow, a convenient sample. Go beyond that. It just put up leaflets on campus, recruit a diverse set of students, but get beyond this campus, put up leaflets in the streets.

[00:28:32] And quite often, if you look around, if in Psychology, of course, they're very aware of that and they might have a participant pool of people outside your university campus who are happy to join in experiments, these might be pensioners. You might have that gender balanced then for example, on Psychology students, we get a more female sample. And so it's not that hard to get more diversity in your participants. It might be a bit harder. If you're in, I don't know, in Europe, you might struggle for example, to get a people of color, for example, to join in your experiments. But I think with a bit of effort, we can get more diversity in our results and we should.

[00:29:11] **Christoph:** For many participants, it will be the first time they have a chance to interact with an actual robot. This results in an overlay enthusiastic response often described as the novelty effect. It is fun to interact with the robot for the first 30 minutes. What happens afterwards?

[00:29:27] **Tony:** Yeah, the first encounter with the robot, we've got this novelty effects, robot is a lots of fun, but it tends to wear off, that's not really a problem. All enthusiasm for anything tends to wear off after a while. So even if you meet people, the first 30 minutes might be great. And then after you go yeah, okay. Done. And the enthusiasm kind of mild just the little, the trouble is. But in human robot interaction, we often stop after the novelty effect has worn off. So we measure those first 20, 30 minutes of interaction report on that. But we don't look beyond that.

[00:30:01] We don't look at, subsequent encounters where indeed people get used to the robot and get used to it its limitations sometimes even get bored with the robot. And actually those moments are perhaps even more interesting than those initial minutes where the novelty effect is still at play.

[00:30:16] Another aspect that we have to deal with is the Hawthorne effect. And that's where, it's named after an experiment from almost a hundred years ago, where people at the Hawthorne factory trying to measure what increased productivity, and they tried all kinds of things to reduce the lighting, increase the lighting.

[00:30:33] Give people more breaks, et cetera. And it seems that whatever they did increased productivity in that factory. And in the end, it turned that it was the experimenters walking around with clipboards that caused the productivity to go up. So it's the observer has an impact on the results of the experiment.

[00:30:53] And we see it as well with robots, the moment people step into your lab, and they know that they're being observed because he had to sign a form saying that they agreed to give up that data. At the moment they know they're being observed. You've got this Hawthorne effects coming into play. Your results will be influenced just by the mere presence. And by the fact that you're running a study.

[00:31:15] **Christoph:** I suppose this is particularly important for longer term studies, but it seems like many researchers shy away from them. Why is that?

[00:31:26] **Tony:** People shy away from that, just because they're hard to do. I think it's not that we don't want to have those results, but it's such a commitment from the research team, but also from the people who interact with your robots, that long-term studies are, few and far between. Imagined that you need to go to a school and you want to spend, I don't know, three months in a school, that means that you need to gain the trust of school management, the teachers, the parents for you to be in that school to run your robot experiments there.

[00:31:55] And three months of collecting data is hard slog. It really is. If you can get the publications by just running a 20 minute experiments. Why would you want to run long-term experiments? But I think that long-term experiments are incredibly important. They tell us so much more than a short experiment does.

[00:32:13] **Christoph:** If planning and conducting an HRI study has not challenged you enough. Try analyzing one. Tony, what is the problem with the use of statistical methods in our studies?

[00:32:31] **Tony:** Oh, my. Statistical methods, we've caught onto the fact that we need to be really clued up about our statistics. And so people will, collect quantitative data and then move to analyze that quantitative data. But in recent years, we've realized that this really isn't always the best way of getting at your insights and results.

[00:32:54] Yeah. So we are a bit plagued by this null hypothesis testing problem where you put out a hypothesis and then you look to confirm or falsify it. And you used that by your, or you do that by collecting data and using statistics. And we've agreed over the last decades that's a small number, the p number, which gives you the probability of actually seeing an effect that isn't really that. This, p number is the, has almost a, kind of a religious status in science.

[00:33:30] Yeah. And so we all want the p values to be below a certain pre-agreed number in human robot interaction that's often 0.05. And what then happens is of course, that people set up their entire study around lowering that p value. So they might do p hacking, meaning that they will keep going at collecting data until the p value drops below that 0.05.

[00:33:59] And that is of course bad practice. And so we need to move away from that. There has been, a whole upheaval in psychology and many other fields where we show that as scientists, as humans really we're just to enthralled by that p value. And by null hypothesis testing, and there are good alternatives right now to move away from null hypothesis testing.

[00:34:22] **Christoph:** If Frequentist statistics is so problematic, what are the alternatives?

[00:34:27] **Tony:** Could just report more than just your p values. So for example, I don't know, effect size is really interesting. So it's not really a bad, have you found something that is statistically significant. In the end, we want to know, Hey, does it work or not? And how big is that difference? And that is so much more interesting. And so an effect size, for example, if you report that is of so much more value to the scientific community, but we can also move to other ways of analyzing the data. So Bayesian statistics is quite new and quite esoterical really, but it does away with the problems that we have with null hypothesis testing.

[00:35:08] **Christoph:** I collaborated with a statistician to apply Bayesian statistics, to some of my studies. In the reviews we received the reviewers often admit that they are incompetent to judge the correctness of our Bayesian statistics. Many current research are unfamiliar with this statistical approach. Hence the paper can be rejected just because no suitable or competent reviewer could be found. How do we deal with papers getting rejected because the reviewers didn't understand your statistics?

[00:35:37] **Tony:** I think I've seen the opposite really where the statistics are just so mind boggling, that reviews just going "Yeah that looks really complicated. That'll do". And the paper gets in even though the results aren't that fantastic. The statistics and the tests are so esoteric, uninterpretable so for someone who'not a, an expert in Bayesian statistics that they just let things through without checking. Yes. It is a problem. I'll admit that the Bayesian statistics that you'll find in publications of my team, I didn't do. My research collaborators did those and I can look at the final results and go yeah, that seems to check out.

[00:36:18] Don't ask me what they're really about. I have to trust my research team blindly to get the statistics, perhaps we need to be more clued up about Bayesian statistics. Perhaps we need to educate ourselves. And I think when better software becomes available, software, that is easy to use, software that we really understand, what goes in and what comes out.

[00:36:38] Then that problem is going to go away. But for the moment indeed as you say, we've got this problem where no one really can assess the quality and the contribution of Bayesian statistics in fields.

[00:36:49] **Christoph:** I've recently started to embrace another safeguard against the abusive use of statistics. I asked my students to register their experiments prior to their execution. They need to describe the exact nature of the experiment and their prediction of the results. This way they cannot go fishing for significance, meaning they run all possible statistics and hope that one of them will come up with a statistical significant difference. Do you, pre-register your studies?

[00:37:16] **Tony:** We have started doing so. So the first one where we, that this was a large scale study where we wanted to see if children could learn a second language from a robot. And it was, it took us a long time to set up and we had over 200 kids take part in the experiments and we wrote then a hypotheses and a methodology.

[00:37:40] And indeed preregistered that. So it gets locked away in a fault and you can't change it. It turned out to be a golden move, really, because as we did this, it increased trust in what we did in the end, we found the no results. So there was no difference between the conditions that we had in our study, but because we preregistered this, it somehow convinced the reviewers who looked at our paper of our integrity, of our intent of doing honest scientific research and the paper went on to win a Best Paper award at HRI.

[00:38:13] **Christoph:** Is this your 2018 paper, "Social psychology and human-robot interaction and uneasy marriage?"

[00:38:18] **Tony:** No, it's not that one. No.

[00:38:21] **Christoph:** No. Okay. Just checking because in that paper, you report on an attempt to replicate a previous study. And when you ran the statistics, you received different results. Why is that?

[00:38:33] **Tony:** Yes. It was interesting. Wasn't it? So it was just a time when the Pepper robot came out and we just had one in the lab and we thought, "Oh, go on, we need a quick results with this new robot that we can submit to conference". And so we went to check some papers on social influence from social psychology and thought this is going to be super easy to replicate with this robot.

[00:38:57] Implemented the whole thing, went to the library and invited the people and looked at the results. There was nothing. Yeah. We really thought that people would somehow their performance on a task would change due to the social presence of this robot. Absolutely nothing. We changed the task again, nothing. We had a different task, nothing.

[00:39:17] It was just incredibly disheartening to see that results that were textbook in social psychology couldn't be replicated with a robot. We still don't quite know what went wrong there. Possibly, it was again, the Hawthorne effects showing up in every condition people felt observed. If not by the robot, they might felt, they might feel observed by us who were leading the experiments.

[00:39:43] And so we didn't see anything. And between different conditions, there was just no difference at all. I don't know. It was just so weird to invest all that time and effort in running that experiment, then having the entire lab take part was really six, seven people taking part in that whole experiment for then to have a no result.

[00:40:02] **Christoph:** Did you preregister that particular study?

[00:40:05] **Tony:** I can't, I don't know. I don't think so. No, I don't think so. It had to move so fast that we didn't preregister.

[00:40:12] **Christoph:** Is this a problem just for this specific experiment that you ran or is this a bigger problem?

[00:40:18] **Tony:** It's, I think it's a bigger problem. We don't really do replication studies on HRI, and it's such a problem that last year at the HRI conference, we had a special track that only took replication studies. So people who dared to take a result that was established in the literature. And rerun that study and see if actually that result checks out, they could submit to a separate track, which I thought was a brilliant thing.

[00:40:45] **Christoph:** And what happend?

[00:40:46] **Tony:** It's a mixed picture. Let's say so some results are, can be replicated, but sometimes even results that are quite. I don't know where you're going like this should check out, doesn't work at all. And it's down to the smallest things. I can give you an example from a study that we run. So that was a study from Yale University team of, Brian Scassellati and they had a result where they could show that people became faster at solving puzzles in the presence of a robot. So that kind of robot encouraging them, giving them hints, make them much faster than if they just got those hints and encouragement from a computer screen. Now we thought Hey, that's an easy one to replicate. So we brought that to Britain at the time.

[00:41:29] So at the University of Plymouth, we decided to re-implement that experiment and we didn't find a single thing. Yeah. And then if you dig a little deeper, what we found was that the the population used in the study in Yale was actually Yale Computer Science students. So very clever kids who were quite sensitive to that.

[00:41:50] So there were just that the right edge to be stimulated by the robot. So you see the result in Yale, move it to Britain, where we were dealing with a population that was more diverse and not quite as talented as Yale students and suddenly those results disappeared. So it shows that result that you take at face value that robots, that encourage you have this power to make you faster at difficult task that this, if you wouldn't replicate that you would think that this result transfers to every situation and through our replication study, we realized that it doesn't, that there are situations where you don't see that effect.

[00:42:27] **Christoph:** One of the main criteria for publishing studies in the best journals and conferences is novelty. Replication studies have a much lower esteem and hence only few bother to replicate even their own studies. There's a clear bias on what studies get published and which one does not. This includes the difficulties of publishing studies that did not show a significant difference. What are consequences of such a publication bias?

[00:42:55] Tony: Huge consequences because replication isn't really welcomed. So I think the HRI conference is the only one since this year. Explicitly send out this message. If replication is valuable, send this your replication studies. But otherwise they can get swept under the rug, really. It is a problem together with studies that have a null result.

[00:43:18] These studies just end up in people's drawers. So you've got this thing where only those studies that are successful, that you know have a significant result. The report get published and the rest is just, just disappears.

[00:43:35] Christoph: And what kind of consequences does that have for the development of scientific knowledge?

[00:43:39] Tony: Consequence really is that a lot of mistakes that we make get made over and over again. So if we don't learn about null results, if you don't learn about what works and what doesn't work specifically, what doesn't work, then we will try and reinvent the wheel over and over again. And that's just. Yeah, w we shouldn't do that.

[00:43:57] We should have a way of sharing all that knowledge with each other. But at the moment, we still have to go pass the gates of a conference or a journal before we can disseminate our results. It sets us back. I think it really isn't, what we're doing ourselves a disservice as a field.

[00:44:14] Christoph: In my 2010 paper entitled "The All-in Publication Policy", I argued that all papers that are written eventually get published. It is only a question of where and when. The peer review process is therefor only a sorting mechanism and not a filter. Still I'm experiencing enormous pressure to publish only in venues that are being indexed by Scopus, which is an indexing company owned by Elsevier. One of the main reasons is that my university is heavily concerned about their worldwide ranking, such as through the QS rankings, which use the data from Scopus. How do you decide where to publish your studies?

[00:44:56] Tony: And your right, you know, in the end, the money decides everything. So your university will have to stick to the framework set by your government on how it receives funding. I'm affiliated with two universities, the one in Belgium, Ghent University. It has a similar system like New Zealand. So we look at Web of Science.

[00:45:15] And if the journal is ranked in the Web of Science, then a publication in that journal is going to attract funding from the government. If not, it really is a dud and it doesn't get financially rewarded. But I also work in the United Kingdom at the University of Plymouth and there they have a better system.

[00:45:33] So every seven years, the government. Once every research scientists to report their four most influential outputs, how these could be publications, conference papers, I don't know, journal articles doesn't matter could be a book. It could be a blog post even could be data. It could be an algorithm, but only four you need to report.

[00:45:56] And you need to explain, in just a hundred words, why you think that output is worth sharing? And that gets assessed by a panel and then gets given a star rating. Yes. So from one star to four star, four stars, really meaning that, the international community has picked up your results and it really influences scholars on a global scale.

[00:46:18] That is, I think a much fairer way of doing phase of things. It isn't knowledge isn't guarded then by the. Yeah, the Elsevier is of the Web of Science of the world. Instead, it's much more in the hands of the scientist. And so I prefer to UK system.

[00:46:48] **Christoph:** During your visit in New Zealand in 2016, did you have a chance to visit Whakaari White Island

[00:46:54] **Tony:** I didn't no, I toured the south Island, but I never made it to a Whakaari White Island.

[00:46:59] **Christoph:** Last year, the volcano on the Island erupted and killed 22 and wounded another 25. This is a true tragedy. And you might wonder why I bring up this event?

[00:47:10] **Tony:** I am very much so. Was there a robot harmed as well?

[00:47:13] **Christoph:** No robots were harmed and none of them were even used in the rescue. But this week, the WorkSafe government agency charged several organizations, including GNS Science, which are responsible for alerts of a volcanic activity at the Island. The problem is that scientists are now being charged in court for their communication, similar to what happened in the aftermath of the earthquake disaster in Italy in 2009.

[00:47:41] The problem is that when scientists become legally responsible for their communication, we would need to become very careful and selective when we talk to the public. Most scientists would probably not speak up if this could get them in jail. Here comes my question. How can scientists continue to be a critical conscious of the society if they lose their academic freedom to express their views?

[00:48:04] **Tony:** I'm just speechless. So how come as a scientist, you can be held responsible for what you say. That shouldn't be the case, right?

[00:48:13] **Christoph:** Well, Tony, in your own, what you just told me in your television show. If you gave people the impression that a robot has a certain ability, and let's say certain people would make certain decisions based on that. And maybe they would lose money or some other form of damage. And they would come back and say yes, but we it

[00:48:32] **Tony:** It is true. Yep. Absolutely.

[00:48:34] **Christoph:** and you and charge you in court. They could do that.

[00:48:37] **Tony:** I agree. This is close to home that I realized, yes. What if people make investments based on what I say or what if people indeed buy a robot to provide therapy for their, I don't know, mentally-ill child or something, then I don't know someone gets hurt.

[00:48:51] So I don't know, could I be held responsible? I, it's a really difficult one. As a scientist, we have a responsibility towards society, and I think that we should be very careful in our communication and should give all the facts and the trouble is that quite often we're being edited. If you give an interview you to a newspaper or TV program or something you're being edited to you're being reduced to a soundbite.

[00:49:15] I don't think it's necessarily the scientists. I think the majority of the very large majority of scientists I know are very careful in their communication and very subtle as well. And if you produce, if you interview them, they're very happy to engage with you and explain more about what is happening that are the exceptions, of course, but these are really exceptions.

[00:49:34] But I'm more worried about the media really, especially in my case, in our field in Human-Robot Interaction, the media has, they already know which story they want to tell. They just need a scientists to give that. Their report or their article, just to, this is a little extra je ne sais quoi.

[00:49:50] Yeah. There's this little extra something this thin layer of respectability, because there's a scientist that has been interviewed, but they already have it as a story. And to that, I really object. I think as scientists, we should, guard, the message that appears in the media and we should own that. Not the media.

[00:50:08] **Christoph:** In this example of earthquakes and volcanoes, there's a component of uncertainty in it. Predicting an earthquake predicting a volcano eruption is very difficult and very dramatic. And people lose their lives. So it's an extreme example, I should say. But nevertheless, the example that you encountered, where your robot was presented in television to have certain abilities might mislead people in believing that the robot has certain abilities, which it doesn't.

[00:50:37] And I certainly take your point that the media plays a role into this. They want to show robots in a certain way. They want to show us a certain future or a certain possibility for a future. But I have to admit that I very often get very worried about it because we're constantly selling these fantasies about robots can and cannot do.

[00:50:56] And that influences not only our participants in our experiments, but it also influences the people in charge of funding. And we have to tell them ever increasing fantastic stories about what our robots can do in the future. And, I think that is quite problematic. How do you deal with making funding proposals and the promises in it?

[00:51:20] **Tony:** Everyone is in on it. And funding proposals do exaggerate things. They, I think people who write those proposals and people who judge those proposals, they're in on the game, they realize that you're making certain bold claims. And I think in funding proposals, that is to certain extent allowed because those proposals are pretty private.

[00:51:40] No one gets to see those funding proposals, except a handful of selected people. Oh, I think the real danger is when we speak to the public and start making claims of, I don't know how wonderful new technology is going to be in what we all can do. We're seeing this, with artificial intelligence everyone now thinks that artificial intelligence is this amazing technology that can solve every problem.

[00:52:03] Now it is an amazing technology. But it cannot solve every problem. It has serious limitations, which we're starting to encounter now and quite serious as well. There's enthusiasm for new technology. That the way in which we embrace new technology is sometimes such a rapid pace that we don't think about the consequences before it's too late and we've done it time and time again, combustion engines ,I don't know, nuclear. Nuclear weapons. I don't know global warming is all because we embraced technology before really thinking through the consequences. And I hope that we don't do the same with artificial intelligence and robotics.

[00:52:39] **Christoph:** This is a good ending statement, I think, and I hope that our listeners benefit from it. I hope that this discussion helps you to improve your next HRI study so that it won't be your last. Remember nothing worthwhile is ever easy. Thank you Tony for joining this episode.

[00:52:58] **Tony:** It's an absolute pleasure being here. Thanks.