Barefoot Innovation Podcast with Nell Watson, AI Ethics Maestro, IEEE Standards Association

***Note that transcripts may sometimes contain errors and that transcript timing notations do not match the posted podcast**

Jo Ann Barefoot: I have been very excited about today's show because we're going to be talking about all things AI today. My guest is Nell Watson, who is the author of a new book called Taming the Machine, and is an IEEE AI ethics maestro. Nell, welcome to the show today.

Nell Watson: Thank you. It's a great pleasure to join today.

Jo Ann Barefoot: I had the great pleasure this weekend of reading the preliminary version of your book, and I could talk with you for hours. There are so many interesting and enlightening things in your book and I highly recommend it to everyone. Again, it's called Taming the Machine. But before we jump into the AI topics, tell us about yourself. What's your own background and what brought you to this subject matter?

Nell Watson: Sure. Well, I have a background in machine vision and that's all about teaching computers to see the world, to perceive pictures, videos, and indeed 3D environments like a robot and navigating inside a building.

I was working on a very difficult problem. It was relating to a startup that I had at the time, which is still going, which was working on making it really easy to measure your body just using a camera. So one picture from the front, one picture from the side, a lot of machine vision magic and we could reconstruct your body in 3D. But there was a real problem and that was cutting the person out of the background because you don't want to be measuring wallpaper behind somebody, right? You want to get it pixel perfect, just their body and nothing else.

And we handcrafted these algorithms to try and cut that person out and the head would fit, but the crotch would break, or one of the arms would fit, the other one would not. It was a real nightmare to try to program this by hand. And to be quite honest, we were kind of stumped at this.

However, this was around about 2011, 2012, and it was just at the point where we were starting to see the emergence of, revolutionary for the time, AI techniques in deep neural networks, deep learning, and convolutional neural networks in machine vision. Suddenly we had systems that were able to understand objects within a scene, quite complex understanding of what was going on and to recognize the difference between a cloud and a teddy bear, for example.

So we actually took several hundred examples of what we wanted, basically manually cutting the person out in photo editing software, fed it into the system, a little bit of crunching for an hour or two, and suddenly that system worked

flawlessly every single time. It was a magical moment and it really taught me just the power of machine intelligence and indeed where we were going with this new wave of working with data and working with machines.

And of course in the years since then, we have seen a progressive acceleration of these technologies. They've gone from interesting in very niche, lab experiment areas towards of course the age of generative AI where any one of us can work with a machine and get it to do our bidding. Using nothing more than a natural language request, we can receive all kinds of incredible content.

And I suppose in witnessing the gradual growth of AI technology, I became quite an evangelist having seen myself just how transformative it was. But as the years went on, I also of course started to notice some of the issues of these technologies, that they were a bit of a dual-edged sword and they could be applied to do incredible things in the world for good. But also, sometimes bad actors could use them in various ways and sometimes well-meaning people could use them in a way that whilst well-intentioned could lead to various misapprehensions in systems and that could lead to unfair and unpleasant outcomes for some people.

So I wanted to do something to help that situation, and I felt a bit of responsibility having contributed to some of these capabilities with patents in machine vision techniques, et cetera. And so, I've been working now for almost about 10 years on developing new standards, certifications, and professional credentials for responsible usage of AI, and ensuring that we can improve the transparency of systems, or accountability so we know if something went wrong, why, and how to prevent that in future, sustaining privacy, and reducing the issues of various forms of bias that shouldn't be in the system.

Jo Ann Barefoot: Great. I'm looking forward to drilling down on all of those topics. Before we jump in, paint the picture for us a little bit of how you think AI is going to change our lives beyond what everybody is saying on the surface. How big are these changes going to be? Where will we notice it the most in say, five to 10 years from now?

Nell Watson: Well, it's strange to think that at the time of this recording, it's only been about 15 months since the release of ChatGPT, which was such a Sputnik moment for the public consciousness in realizing just how far we had come with AI. And of course it had been bubbling under the surface for a few years, but hadn't gone noticed by many people. And I think we're actually at a similar period with a new phase of AI, these agentic models.

So a lot of us are starting to realize that the generative AI systems sometimes hallucinate things. They confabulate, they make up things that aren't actually real when they don't have that much training data to work with, and sometimes human beings do the same things. But we're learning actually that we can steer

the thinking of these generative models using something called scaffolding, which is like using little programs to help to guide the thinking of these systems.

And so actually we can help them to reason better. We can help them to check their mistakes. We can help them to use tools to bring in new information or to perform functions in the world. And all of this enables these systems to be much more of an independent, reasonable thinker. So they're not just so much an intelligent tool, but they're actually capable of significant autonomy.

These systems can be given a broad mission and from that they can generate all kinds of sub goals, which would fulfill that mission, a plan of action step by step. And they can even split themselves into multiple pieces, a little bit like Agent Smith in the Matrix, and they can delegate different tasks to different sub-versions of themselves. And so they can attack a problem from multiple different directions at once.

And so these age agentic models are going to be a very powerful, forthcoming, imminent next wave in AI. And it's going to make these systems able to act as a concierge to be able to solve all kinds of complex problems, whether it's organizing lunch with friends or whether it's putting an event together or deciding the ideal specific travel itinerary to take that accounts for a wide variety of different factors.

So these systems are going to be much more capable, but there's also going to be a great deal more challenges to come with them. Because these systems are able to take actions in the real world, they're able to put goals into action, it's very important that we teach them how to fulfill goals properly. If you're organizing a picnic, you don't want the system to decide that a nine-course meal is ideal for that situation. You also don't want the system to decide that giving everyone a wafer and a shot glass of tap water is going to fulfill that remit, either There's a balance of reasonableness that these systems need to be able to learn.

And they need to be able also to learn our values and our preferences as well as of course culture, to be able to read the room and fit in. If the picnic that you're organizing is for a mosque or a synagogue, the system ordering ham sandwiches for everyone is not going to fit in with those values of that community either.

And so there's all kinds of problems, therefore with goal alignment and value alignment, which are moving beyond AI ethics and actually into a new realm called AI safety, which until recently has been largely theoretical or largely science fiction, although a very important area of research. But now this is actually coming into our daily lives and it's something that all of us are going to have to wrestle with to some degree or other.

Jo Ann Barefoot: So I want to come back to the ethics and safety, but let's first go a little bit deeper on the agentive AI. One of the points that you make in your book many

times is that there are uses that are more high stakes where errors are going to really do harm. I like to say if Netflix recommends a movie to me that I don't enjoy, there's not much harm done. I may have wasted a couple of hours. But if my doctor makes a mistake, that's a different thing.

So in financial services, I know our audience is interested in everything about AI, but we are focused on financial services, what is the prospect in your mind that an individual would be able to have an agentive AI to guide her financial life?

And when we think about the tasks, there's a realm of tasks that are about daily financial management, bill paying, budgeting, setting money aside for savings and so on. There's another realm that's about financial advice, which is more complicated, what kind of an investment should you make? How much risk should you take?

And there's another realm that I'm very interested in, which is in financial shopping. Might we have agentive AIs who could help us recognize that a certain financial product had a lot of hidden fees or so-called junk fees, might look good on the surface, but might actually trap people into issues that they don't realize at the front end are there? How realistic is it to imagine that our AI friends might help us with problems like that?

Nell Watson: I think that the picture that you paint here is very realistic, and we are going to see that within say five to 10 years possibly earlier. Machines are going to be helping us to decide what kinds of purchases are best likely to fulfill a certain need, whether that's a beautiful gift or whether that's trying to stretch every penny and get as much out of it as we possibly can.

Indeed, that's probably going to also involve things like investments or trying to find the right kind of investment which balances risk and potential reward for example. We also might use machines to sift through information and to potentially spot an opportunity or some nature of arbitrage that otherwise might go unnoticed on the surface.

Indeed, machines can also help us to create and negotiate contracts with other people. They can actually help us to find win-win solutions and to map out the non-negotiable values of different parties, and therefore help them to understand each other and to get to an equitable solution that everyone is reasonably content with.

We already are quite used to translating between German and Swahili using Google Translate, for example, or DeepL. Pretty soon machines will help us to translate between values and culture to understand each other better. Or indeed perhaps understand why we ourselves might be in a bad mood, like have we eaten today? Did we sleep properly? Maybe that's why we're cranky and maybe

all these other people aren't necessarily as big jerks as we might imagine that they are in this heated moment.

And so I think that machines can really help us with navigating a lot of different complexities in our lives. However, I'll give another caveat that we should be careful because we've seen examples such as the Horizon post office scandal in the United Kingdom whereby an algorithmic system which was designed to detect fraud, wrongfully implicated hundreds of people in having committed fraud.

Dozens of people were sent to jail for years for crimes they didn't commit. They had to sell their houses to pay back a debt that wasn't theirs. Marriages were broken up. Tragically, some people even took their own lives as a result and would never live to be eventually vindicated after many, many years, because this went on for about 15 or 20 years before it finally really got the attention that it deserved.

And unfortunately, this is not an isolated case. We've seen in the Netherlands, the Dutch child benefits scandal whereby there was a disproportionate interest in persons whose first nationality was not Dutch, even if they had since become naturalized citizens. And there were some ugly threats made to people to take away their children and things like that. And of course, such a furor that actually the government of the Netherlands collapsed as a consequence.

And we see the same thing in Australia, the Robodebt scandal in Denmark and in Michigan, tens of thousands of people wrongfully accused and given the third degree basically for issues that honestly they were innocent of fundamentally.

And that's why we should be very, very careful that we have to ensure that there is adequate human oversight of these kinds of systems, that we don't blindly trust what they're telling us, particularly when it involves whether somebody might have access to something like employment or education or housing, and especially of course, where a system might ultimately decide somebody's fate in the judicial system or the medical domain for example.

So I think it's important that we understand the risks of technologies and that we ensure that tech is applied in use cases where there is little risk, and the riskier ones use technologies which are much better understood, have been shaken down, and are easier to explain and to interpret. And so we can more readily understand if the system's doing something that perhaps it shouldn't.

Jo Ann Barefoot: I want to go to these questions of how we make sure that it's being done right, but before we leave the idea of the financial agent, one question I have is, will this be affordable for everyone?

You talk in your book about the potential of machines being able to help with something like therapy and you say it can democratize the access to things that today are mainly readily accessible to people with more wealth. What are the economics of this going to be? Is it going to be possible for everyone to have their own AI agent and will they all be equally good? Or are we going to see that some people have more AI power in their hands than other people do as individuals? Not talking about the government or businesses, but just as people?

Nell Watson: It's an important question because of course technologies potentially can increase inequality in society if some people have access to a remarkably better version of something than others. But I don't think that that's likely to be the case with AI. I think it's actually going to be an enabler of many people, an increaser of general quality between people and between nations and jurisdictions. And that's because already we have many different companies that have different AI systems, and broadly they're all quite comparable. There's some which are maybe better by a nose in a certain direction or a certain domain, but there's a constant shuffling of who's best, and I think that's a good sign.

We're also starting to see very powerful open source models which are developed by and for the community, and it seems that open source is catching up sometimes within less than a year from the world's best proprietary model. The community comes up very rapidly with something that matches or even exceeds that capability. I think that's another very good sign.

Also, AI systems tend to be relatively affordable. We've seen that the price actually comes down quite rapidly following the pattern of Moore's law and such of things getting quite cheaper in terms of price to performance at a geometric rate. And we're seeing a lot of constant optimization of these systems so that they can run on much more modest hardware.

We're now even starting to see these language models, which used to require running on a supercomputer, can now actually run on potentially your own domestic system, though they've been quantized a little bit, which basically means to simplify them slightly. And so their capabilities are maybe not quite so sophisticated, but their 95% is good, but they can run for free on your own system, on your own smartphone or a laptop.

And so I think that this is going to be a powerful tool to democratize capabilities across society, particularly for people who might be on the more vulnerable side of society. Maybe they find navigating modern life to be a challenge. It's not easy to do taxes, it's not easy to navigate a complex public transit system in a foreign city, but machines can help us with this, and I think that's going to be a tremendous boon for a lot of people.

Jo Ann Barefoot: I do think in financial services where we almost always have a marketplace that's asymmetrical between the buyer and the seller in terms of which one understands the product better, you could imagine that if people had an agentive AI and if it could be required that it be optimized for their best interest, not the best interest of a company or something like that, then the marketplace could actually begin to work much more perfectly with people making good choices and providers being rewarded for offering products that don't have say, tricky terms or whatever, because the AI would see through that. So I think it's incredibly promising for consumers.

So going to your point about the fact that these systems can go drastically wrong, and in your book you talk about some of those scenarios, they can develop obsolete data. They can develop obsolete models. I think there's a place where you say that sometimes fine-tuning them can accidentally undo protections that have been built into them before and so on.

As a practical matter, since they're going to know things that we're not going to know, what are the methods? You do lay them out in your book. Talk to us about the methods for overseeing them to be able to be sure that they're not going wrong or doing harm.

Nell Watson: Well, machines do have us at a disadvantage because they're able to infer things about us which are hidden in plain sight. There are AI models today that can make a reasonable guess about somebody's sexuality or even their political affiliation or their intelligence or their character or personality, or whether they're at a vulnerable moment right now because they've had a difficult day.

And potentially, those kinds of things could be used against us. If we're lucky, it might be used to try to sell us something. If we're less lucky, it might be used to try to demoralize us. And so there's a risk of these technologies being used as fifth-generational hybrid warfare whereby instead of attacking a foreign nation with guns and bullets, you do it very subtly and quietly, demoralizing the target society so that eventually it collapses under its own weight. I'm hoping that someday soon we'll see a Geneva Convention against this kind of warfare because if we end up trying to do this to each other in the world, we could be driven crazy by these techniques, and I don't think that's a society that we want to live in.

Mercifully, there are ways in which we can protect our privacy against these systems. There are new forms of encryption which can enable machines to work with encrypted data and make that data useful for training those machines about our preferences, our interests, etc. And yet keep them, to some degree, secret. So we can aggregate our data with other people's, who have a similar ilk or a similar behavior or interest, and yet keep that private to ourselves. Or we can share a very small aspect of ourselves that doesn't give away the fuller picture.

And just as we wouldn't want to do our online banking today, unless we had that padlock symbol in our web browser that tells us we have a secure connection, similarly in the near future, we won't want to deal with an AI system unless we know that our personal information that might be interacting with that system is indeed encrypted in a very secure manner.

So I think that just as the World Wide Web had its Wild West days before it settled down, I think we're in a similar period with AI. I think there will be a bit of a rocky road for a while whilst we adjust to its capabilities and balance the opportunities with the risks. But I think that over time we will learn and we will adapt and we will take AI to a better place.

I liken it to jet air travel in the 1950s in a sense, where we have this very exciting new technology that can take us all sorts of places, but potentially that could also end in tragedy. Some of those tragedies are going to be inevitable. The important thing is that we learn from it, and we adapt, and we take action and make good rules to prevent that from happening in future. And I think if we do that rapidly and cleverly, I think that ultimately we're going to be in a much better place with AI.

Jo Ann Barefoot: Let's talk a little bit more about some of these risks. As we know, the generative AI can not only talk with us and generate text, but also images and the potential for people being deceived. And again, in our realm here, financial fraud and romance scams and catfishing and all these things seem likely to get much, much worse. Is there an opportunity, do you think that if we're going to be attacked with AI by bad actors, that we'll be able to defend ourselves with AI? Will we be able to have a way of seeing through deception by using AI?

Nell Watson: We are at significant risk of being deceived by synthetic media of different kinds such as deep fakes. And these are becoming progressively much more powerful and capable and credible than before. One shouldn't assume that one can always tell that something's been manipulated or that just look at the fingers and that will give it away because things are advancing at such an incredible rate.

And already we can alter somebody's perceived ethnicity or gender or age live in real time or transform them from a sensible person into another, or indeed to translate from English into Spanish, for example, and yet have that person speaking that foreign language that they themselves may not speak themselves with the accent and tone of the original recording.

Indeed, of course, bad actors can clone people's voices from as little as two seconds of speech. So an answer phone recording, "Hey, I'm Bob Smith, sorry I can't take your call right now." That might be enough for somebody's voice to be cloned. And already we've seen examples of heists of $25 million perpetrated because of these voice phishing attacks whereby somebody's voice is cloned,

they happen to be in another nation, they call up a colleague, and the colleague accepts pointblank that they are told to wire money to a certain place, but of course it wasn't that person at all.

So these challenges will thankfully to a certain degree be mitigated by AI. So AI will help us to defend against these kinds of issues to better detect content, which appears to be manipulated. However, it's always going to be an arms race just as it is with computer viruses where people are constantly finding new exploits and these security professionals are constantly patching them or creating new ways of detecting activity that is suspect.

Moreover, there are sometimes challenges with these deep fake detectors in that because of a lack of accurate training data, sometimes for example, persons of color in a piece of media are more likely to give a false positive than other kinds of people, for example. And so sometimes these detectors for deep fakes or detectors for plagiarism can end up having a disproportionate impact on certain demographics.

We see sometimes a lot of wrongful accusations of fraud within the schools or academic systems where people are accused of having plagiarized directly from ChatGPT until somebody puts the Gettysburg Address through the system, and it also gets flagged, for example. So it's very important that we're cautious and that we are always checking or checking mechanisms basically.

Jo Ann Barefoot: Let's talk for a moment about these issues of bias. And again, you have a lot on this in your book. The risk that these systems, and an area that is very prominent for our listeners is credit underwriting. Well, we've moved to a system where we're using more data and therefore we should be becoming more accurate in recognizing risk rather than just using a few limited markers of risk as we tend to do today. But we also know that there is a lot of structural bias in the data that would be learned from. And you see this even more in something like, I think, criminal justice and other areas. How are we going to defend against that? How are we going to keep this from either introducing new bias or exacerbating old bias because it's learned from old data?

Nell Watson: It's a very tricky problem because these biases can come in from a range of different vectors. For example, the data itself might be messy. It might have all kinds of gaps or lacunae in it, and that can be a problem. It can be older data, which perhaps does not reflect current realities.

For example, a lot of trading systems like at hedge funds and things, whenever the pandemic hit, they went haywire because this was a black swan event that they weren't able to map to previous assumptions about the world or economic functions or how seasonality affected air travel, et cetera. That all completely went out the window. And so those systems were stumped at how to interpret that kind of thing.

We can also see a reiteration of previous patterns, if a system learns to detect an older pattern. For example, if a lot of engineers happen to be male or a lot of kindergarten teachers happen to be female, the system may say that a kindergarten teacher necessarily should be female then because that fits the pattern of what's worked in the past, even if that's not necessarily objectively true or indeed an ideal.

So it's very important that we test these systems quite carefully, that we look for ways in which machines could have learned the wrong thing. For example, in the medical world, there was a system which was designed to do risk classification of patients, basically decide how risky a certain procedure would be for that patient.

And it determined that Black Americans required less medical care than white Americans because it had learned a pattern that Black Americans as a group tend to spend less on their medical care for economic reasons. But it had learned the wrong thing and gotten the wrong end of the stick, and therefore reasoned that because Black Americans spend less on care that they need less care, which is of course absolutely not the case and the wrong lesson to be learned from the data.

So that's why we have to be very cautious and to test quite carefully what these systems have learned. And we can do this through things like unit tests, various forms of benchmarks, for example, which can help us to test a system and to test it every time it's updated, every time it learns a new set of information, it's given more information about the world, et cetera. Those tests should be undertaken to make sure the system hasn't drifted from its original intent.

And unfortunately, machines can and do drift sometimes. They may work well at training and then after a month or two they've gone off on a strange tangent and they're no longer working the way that they did. And if somebody isn't carefully monitoring for those issues, then strange phenomena can occur in these models and that may not necessarily be picked up upon. And of course, that can lead to ethical or safety issues down the line.

Jo Ann Barefoot: And do the methodologies already exist for conducting this kind of testing, or is that something that we are building as we go along?

Nell Watson: We've made incredible progress in recent years. We're no longer dealing with abstractions of principles. Principles are great because they're timeless, they can be timeless. The Peelian policing principles are almost 200 years old, and they talk about policing being done with the cooperation and consent of the community. That's a timeless value, and it's still very valid today, even though of course policing and forensics have changed an enormous degree in the last two centuries.

However, it's important that we have actionable rubric, that we have benchmarks that we can create a sort of an acid test of a different system or indeed the organization behind it. And that's what we do have today. We have these standards and certifications that we can directly apply.

And that's going to be a very important part of the branding of various organizations because they can demonstrate, hey, our system is quantifiably better than the competitor in these areas. And in a world where people are looking to embrace AI, but also very concerned of potential drawbacks or even scandals, being able to measurably demonstrate that you have de-risked your product is going to be a huge asset to any firm.

Jo Ann Barefoot: We know that people are experimenting with AI everywhere now, companies and governments and individuals. Do you have a sense of whether these best practices are being widely used or haven't they caught up yet with the rapid explosion of experimentation?

Nell Watson: There is a slower diffusion when it comes to adopting new best practices and understanding safer guidelines for using systems, and indeed for generally just understanding the potential risks out there. I really recommend that people have a look at the AI Incident Database. It's a wonderful way of horizon scanning for issues out there in the real world. These use cases of AI having gone wrong or having done something that people didn't expect. And indeed often in those are nuggets of wisdom of how to avoid such pitfalls or how to mitigate them should they potentially occur. I think that's an important place for people to begin.

Jo Ann Barefoot: I know we're running short on time. Talk to us a little bit about your other advice. Your book is set up with each chapter having a good recap of the key things that people should take away and should do. Beyond learning, what is the advice that you have today for say, businesses and government who are trying to build an organization that will do well with AI, that will be able to follow best practices on ethics and safety and so on? What should they be starting to put in place today and then build out over time?

Nell Watson: Well, the most important thing in terms of using AI is data. Data is the oil upon which AI is operated fundamentally. And to use that data, it needs to be in a form that we can readily apply. So it can't just be in papers in a file cabinet, we have to digitize it somehow. We have to verify and validate that that data is still useful, is still applicable, and doesn't have a whole bunch of errors in it.

And we need to put that data in a place that people have access to it, that the people can pull it into various systems and make good use of it. And that's all basic data science, but that in and of itself is already a huge step forward for many organizations, and it's going to be an enormous advantage in and of itself. And it's also going to prepare those organizations to embrace AI.

Now, when it comes to selecting an AI system to test, I would consider to understand the appetite for risk in that organization. Some sectors such as, for example, the energy sector tend to be more conservative by nature, tend to be more cautious, which is a good thing when you're dealing with potentially nuclear reactors or pipelines that could burst. Other organizations are more customer-facing, more about entertainment or things like that, that have little real world consequence to them.

So understanding the appetite for risk will inform what kinds of technologies we can use and in what use cases. And if we're working with, for example, maybe financial information, that tends to be on the riskier side and we should be cautious with what technologies we use. We should try to use technologies which have been established for a while and which are less likely to produce risks because we can understand those systems better. And indeed we can demonstrate to regulators why a decision was made and in a way that also makes sense to the customer.

I think that we should not give up on AI. We shouldn't throw our hands up and despair that it's all so difficult, and there are so many pitfalls. I think that we now have the tools to begin to use AI in a safe and responsible way, not just as an experiment, but indeed beginning to really weave these things in with our professional workflows and indeed customer-facing technologies.

I think that generative AI and now soon agentic AI is in many ways the closest thing to magic in our world today, and I think that we can do wonderful things with it. I think that we can create a lot of peace in the world through these technologies by understanding each other better. We can map various forms of externalities in the world, different effects like pollution or indeed kind things like protecting commonses.

And I think that that's in many ways the missing expansion pack for capitalism, that greater understanding of those externalities and how our actions and behavior can affect other people in negative or indeed positive ways. And I think that AI is going to help us to understand this and indeed help us to charge or to reward people who put things out into the world and indeed help us to steer towards those better outcomes. I think that after we ride the Rocky Rapids for a little while in understanding AI and creating good rules for it, I think that's the world that we will get to in the mid-future.

Jo Ann Barefoot: You start your book with a quote from Jaan Tallinn, which is, "Building advanced AI is like launching a rocket. The first challenge is to maximize acceleration, but once it starts picking up speed, you also need to focus on steering." I thought that was a wonderful way of encapsulating the message. Is there anything we haven't talked about that you want to touch on?

Nell Watson: Not in particular.

Jo Ann Barefoot: All right. Well, as I say, I could talk to you for hours. So the book is Taming the Machine, and where can people get information about you and your work?

Nell Watson: Sure. I have a little website at nellwatson.com. That's Nell, November, Echo, Lima, Lima, N-E-L-L, Watson dot com, and my book is showcased at tamingthemachine.com.

Jo Ann Barefoot: Wonderful. Well, Nell Watson, thank you so much for being our guest today. It's been absolutely fascinating talking with you.

Nell Watson: Thank you so much. It's been a great pleasure.