AI Ignition

Ignite your AI curiosity with Stuart Russell

How do we make AI systems that are beneficial for humans?

From Deloitte's AI Institute, this is AI Ignition—a monthly chat about the human side of artificial intelligence with your host, Beena Ammanath. We will take a deep dive into the past, present, and future of AI, machine learning, neural networks, and other cutting-edge technologies. Here is your host, Beena.

Beena Ammanath (Beena): Hello, my name is Beena Ammanath and I am the Executive Director for the Deloitte AI Institute. Thank you for joining us on the AI ignition show today. Our episode today features Stuart Russell, Professor of Computer Science at University of California, Berkeley, and he is also the director at the Center For human compatible AI. Stuart, thank you so much for joining us on the show today. I have been looking forward to speaking with you today and taking your thoughts to our audience. So, first question, how are you doing? How have you been staying busy in the past few months?

Stuart Russell (Stuart): So, I have to confess that it hasn't been a very difficult period for us. Both of us are able to work from home and we have been able to spend time at a house at the beach. So my son has been doing a lot of surfing, I have learned to make marmalade and a few other things, and I haven't missed spending time in airports. Now, when I look back at my previous life is like why was I spending all that time in airports and airplanes? It's not necessary.

Beena: Great to hear. I have to ask you this question. I was reading one of your interviews and you spoke about how you had a great revelation when you were on sabbatical in Paris, about how do we think about the consequences of AI beyond value creation and I love that. Have you had any great revelations or deep thoughts in these past 15-16 months?

Stuart: That's a very interesting question. I think I have evolved my thinking, I would say refined. So, the big change that happened back in 2013 or so when I was in Paris was essentially this realization that the way we have been thinking about AI was all wrong and the particular impetus for that thought was I was on the metro, at the subway, and listening to a particular piece of music because I was in a chorus, and I

had to learn this music and it was Barber's Agnus Dei, which is an incredible piece of choral music. It starts out with a single note that goes on for 28 beats, and it's just the whole thing is sublime. I was thinking that we spend too much time preparing ourselves to have a better future instead of living in the moment and this kind of quality of human experience is really what it's all about and preparing for a higher quality feature experience is not so valuable in itself. And then it occurred to me that, of course, if we were to build AI systems to help people, they probably would have no clue about this experience and what really matters to people. It went from there very quickly to the idea that we had got the whole way of doing AI, the standard model, as I call it completely wrong. The standard model says write down the objective and then the AI system optimizes it for you. Control theorists and economists, they use the same basic method. Control theorists write down the cost function, figure out how to minimize it and economists say write down the utility function, write down the welfare function, write down the quarterly profit stream and maximize it. You can't do that because we don't know how to write it down. So we are going to have to build AI systems that know that they don't know what the objective is and that's a completely different ball game.

The way I have been thinking about it over the last year, the refinement of that is to say that obviously we should build the machines and deploy them, the machines that are rational for us to deploy, and given that we can't write down the objective, we have to deploy machines that reflect our own uncertainty about what the objective is. So, if you start from the idea that you should deploy machines that are rational to deploy, you actually can derive the three principles that I have in the book, actually is a consequence of just that straightforward idea.

Beena: Can you walk us through an example of what you mean by how AI is currently designed or trained whereby setting the objective upfront? How does that translate into the real world?

Stuart: Now imagine running a company or looking after an old person in their house, you can only begin to imagine how hard it's going to be to specify these objectives. When you look at social media, for example, they use reinforcement learning. They have an objective which they specify for the algorithm, which might be sort of like how many clicks can you generate and those algorithms they have more control over what people see and watch, the cognitive input of the world's population, they are more control than any dictator in history ever had. What do they do? Well, they do reinforcement learning. In reinforcement learning, the algorithm learns to come up with a sequence of actions that maximizes the long-term sum of rewards. How do you do that? Well, what do the actions do? The actions present information to the user articles or YouTube videos or whatever it might be, and the state of the world that the algorithm is learning how to change is your brain. So, they are learning to manipulate people so that in the long run those people become more predictable consumers of content that they will then send them. So, they don't just learn what you want, they change you into a different kind of person than you were before. So, that they can generate a larger stream of rewards.

Beena: To a large extent, to think of all the possible rules, the objective, like fine tuning that objective is hard, almost impossible because you cannot possibly think of all the scenarios that might come up. I have listened to your Ted Talk on creating safer AI and you shared three principles. Can you share some of the thinking behind it and tell our audience what those three principles are as they think about using AI within their organizations.

Stuart: Sure, so I think the key idea is whenever you design an AI application that you are thinking about deploying, first of all, try to figure out what the objective is, what is it that we would like to have happen

when this AI application is deployed, including thinking about the externalities, not just can I generate profit from it, but what other changes can this bring about in the outside world, what are the implications of those. The second stage is then to say, well when you actually try to do that, you realize that you can't do it completely. You may be able to have some partial idea about what matters and what needs to be paid attention to in the objective function. But then, as with the example of driving, there are always going to be other things that you don't know and you may find out later by seeing what happens. Oh, we forgot to put in the thing about passenger comfort or not freaking out the passenger. So, the way the three principles work is to say that the first principle is that we want AI systems that are beneficial to humans, essentially given that humans may have preferences about how the future unfolds, the AI system should act in such a way to further the preferences of humans about the future. So, it's a very general formulation of what we mean by beneficial AI. The second principle is that the AI system does not know what those preferences are. It may have some partial information up-front. The third principle says that the evidence that the AI system can use to learn about human preferences is human behavior. Our preferences about the future that we want are partially and noisily revealed through all the choices that we make.

Those three things together, those three principles actually end up defining a new kind of mathematical problem. So traditionally in AI, there is a whole sort of hierarchy of problems getting more complicated, but they are all problems in which a single agent, the AI system, is going to optimize an objective of some kind. With this framework, we get a different kind of problem, one in which the machine and the human are both participants because the human is the one that has the preferences and the machine is the one that's supposed to be helping the human, but because the machine doesn't know what the preferences are, there has to be some flow of information at runtime, so to speak, from the human to the machine about what the human wants. So both the human and the machine are participants, in economic language we call this a game because it’s a game meaning a decision problem with at least two participants, and when you solve that game, you can then write down mathematically particular games or particular kinds of settings and you can solve them, you can calculate what is the solution for the human, what is the solution for the machine, what does it look like and what you find is that with this formulation, the machine becomes deferential to the human. In the classical formulation, where the machine believes it knows the true objective, then it doesn't defer to the human at all, it simply carries out whatever action it thinks is going to achieve the objective. Even if the human is jumping up and down saying stop, you are destroying the world, the machine says I don't care, I am doing the objective and the objective is true and I am maximizing it. End of subject. So you get this kind of single-minded pathological behavior. So, the new framework, the machine defers to the human.

If the human says stop, that's conveying preference information. It's saying whatever it is you are doing, I don't like it. So, now the machine updates its understanding of human preferences and says okay, I don't want to do that because you don't like it. The first principle says machine has to be beneficial. So, it will defer, it will allow itself to be switched off and this is sort of the core of retaining human control over machines is can you switch it off. In the new framework, the machine welcomes being switched off because that way it avoids doing whatever it was that the human didn't want it to do and it doesn't know why the human doesn't want it to do that because it doesn't know but it wants to avoid it whatever it is. I think that as we learn more about this new mathematical framework, I think we will find that in fact this is the AI we should have been doing all along. These are the products that people will want because they behave in ways that we would like AI systems to behave, and they will adapt to our

individual preferences while also respecting the sort of societal preferences, not acting in an antisocial way and they will avoid making these sort of mistakes. These kinds of systems would avoid these sorts of mistakes because when they know that they don't know the human preference or the human value that might be attached to the cat, they won't construct a plan that involves changing the state of the cat because they don't know whether that's changing the state of the cat is desirable or undesirable for us, and so they will ask permission. They will say, is it okay if I cook the cat for dinner and you can say no please don't, and let me update you about why cats are valuable. Similarly in the global economy, economists are familiar with this idea of externalities. In some ways, corporations that pursue profit while creating massive externalities for the rest of the world are just like the robot that cooks the cat. They are pursuing their objective and assuming that anything that isn't part of the objective, they can do whatever they want with and that's an incorrect assumption. You should have the opposite assumption, anything I don't know about, I should not mess with until I have more information.

Beena: But it's also a little bit of human nature as well to find those loopholes and now how do we encode that within a machine that you put in these guardrails. Part of it also the reality, I remember I studied AI, at that time a lot of it was just theory. We were just thinking about it or debating over it and there was no real way to build these powerful algorithms that we see today. So isn't it this part of the phase of growing up and having these epiphanies and then taking a step back and making sure that we build it all in a very thoughtful, mindful way by using certain design principles. Do you think we are at that pause phase now where we have seen AI can create value in the real world, but there are negative impacts, so let's think about what are some of the things we need to do so that it continues to create value, but there is not as much negative impact.

Stuart: I think you are right, there is a gold rush still going on. Participants in a gold rush are notoriously oblivious to side effects, but I think we are seeing really a massive growth in awareness about possible negative effects of AI. Social media is one example, but in kinds of systems that process resumes or filter out resumes in HR applications. In medical care systems that make decisions about triage and/or about admission to intensive care, etc. All of these applications can be seen as having difficulties for the same reason, which is that typically when you train a machine learning algorithm from historical data, which is what happens with many of these decision algorithms, people assume that the objective function should be consistency with the training data, so that's what they optimize, and typically they even measure consistency by the number of correct and incorrect answers. There is an enormous range of costs associated with incorrect decisions. Just in terms of classifying images, the ImageNet library has 20,000 categories. Typically when you talk about a cost function, you should specify what's the cost of misclassifying an A as a B. So that is 400 million entries in the cost matrix and no one has ever written them down. So, that's the first reason why there is uncertainty about what is the cost function for decision errors. The second thing is that consistency with historical data isn't the objective. What's the objective? The objective is, let's say it's a medical care system, to make sure that everyone gets the best possible quality of care and why should that have any connection to consistency with historical data, particularly if the historical data come from a system that incorporates all kinds of societal biases and systematic racism, and you name it. This comes back to the my first piece of advice to companies that want to deploy AI, sit down and think what is the actual objective that you want to achieve. The actual objective is not be consistent with historical training data, no one actually cares about that, they care about delivering high quality health care, ensuring fair and equitable hiring, or whatever function it is that you are trying to automate or augment with AI.

Beena: So, was that the intent behind setting up the center that you started Center for Human Compatible AI. Is it more toward making AI beneficial to all humans and putting in the human lens on everything that we do with AI? How would you define the Center's mission and how has it progressed so far?

Stuart: The center's name, the Center for Human Compatible AI is sort of a bit of a dig at my colleagues because it means AI that's compatible with human existence. So, I am sort of pointing out that the standard model for AI is not really compatible with human existence in the long run. So, our primary goal is to figure out the converse of that, how do we make AI systems that are provably beneficial to humans. I shouldn't have to say provably beneficial AI, I don't have to say airplanes that are designed not to crash. It's just part of what it means to be a good airplane or a good bridge is that it doesn't fall down or crash or whatever, and it should be part of what we mean by good well-designed AI that it's beneficial to humans. That should have been the definition from the beginning, but it wasn't. We can go into historical reasons why the standard model emerged but when you think about it, you just don't want AI systems that are pursuing objectives that are different from being beneficial to humans.

Beena: Yes, so true. I mean everything that we do, we believe in it's about humans with machines. Humans are always going to be there, it is about how can machines augment us or benefit humanity. What are your thoughts about or what have you seen in companies or in different industries where AI has been most beneficial without as much unintended consequences? Which industries are actually building provably beneficial AI? Which industries are actually succeeding at it?

Stuart: I would say at the moment, the technology that the center has been developing is not ready for a large scale deployment. So, as I said we have been able to formulate and solve some of these assistance games and examine the behavior that results and it seems to conform to our qualitative expectations that it does defer to the human. It's interesting to see how the human behaves in these games as well. The human has an incentive actually to teach the robot the human preferences as quickly as possible so that the robot gets up to speed and is helpful. We have seen interesting ways that the human half of this algorithm figures out how to teach the robot. So even in the absence of any communication language that we are building very, very simple mathematical models and simulations, but one of the things it does is it will sort of take the robot by the hand and walk all the way up to the edge of the cliff and then walk away again, and the robot interprets this as don't go over the cliff. So, they actually develop a kind of a communication language between them that allows this preference information to flow. All of this comes out just out of the algorithm solving the game. We are not scripting any of this stuff. We actually didn't expect this kind of behavior and we didn't really understand it when we first saw it, and but now we understand that this is a natural way to teach people not just what to do, but also teach the robot not just what to do, but also what not to do. So. it will be some time, I think before we can package up and sell something equivalent to a Tensor Flow, which is the Google Deep learning software. I think the place that we may be able to have some impact in the near term is actually in social media in helping the social media companies understand how to redesign their systems so that they don't have this manipulative behavior which currently is causing a lot of problems.

Beena: One thing that I have seen, and obviously AI has a very broad definition, but it falls under the AI umbrella. Examples where you are using AI to predict say a machine failure, think of that airplane example, you are predicting when a jet engine might fail so that you can proactively send a field service engineer, technician to fix it, so that there is no unplanned downtime. There is not as much human

interaction or the opportunity for human biases because it's looking at very specifically an engine data and then coming up with the suggestion that there is a 95% chance that this engine will fail in the next 28 hours, for example. How do you think an example like that, what's the lens to look at it from both the provably beneficial to humans and a humble machine perspective? Is that a scenario that would fall in there?

Stuart: So, it's a good example. It's simple because it's not really a sequential example. It's sort of a one shot decision. Do I flag a potential failure or not? So, I should preface this by saying that if the way that diagnostic system works is simply to constantly feed the latest posterior probability of failure, some probability distribution about failure occurring in the next minute, 10 minutes, hour, and so on, if all you are doing is feeding probabilities, then you should simply make that system as accurate and well calibrated as possible. So that's purely perceptual in the sense that it's someone else downstream is actually making a decision using all these probabilities. But if you have a system that's going to turn on a warning light, for example, then that's a real decision, to turn on the warning light or not. How do you make that decision, what's the probability threshold, is it 2% probability of failure in the next 10 minutes let's say, is it 40%-90% and to answer that question, you actually have to figure out, okay, what are the costs of a false alarm, what are the costs of not flagging something that does turn into a failure? So those costs actually have to be thought about and you can't necessarily write them down, and it might vary depending on whether this is a cargo aircraft or a passenger aircraft and is it a cargo aircraft that's frequently flying about populated areas. So, the right decision about what that threshold should be probably requires thinking hard about the true costs of all the possible decisions. The same kind of thing happens in medicine where typically when you get a blood test, they put some arbitrary interval around whether this is normal and then they flag anything that's outside that fixed interval and anything that's inside the interval, you ignore. You say inside the interval, there is no purpose in flagging it, but that's I think a very crude and probably not well calibrated, not well thought out kind of decision. So, I think in in many of these cases, you probably want to have this interactive process of figuring out better of what the human preferences are. In the case of a warning light for a jet engine, you want to have that happen in the design phase. It's probably not going to be feasible at runtime once you are out flying to learn more about the preferences. In the design phase, you start out that design phase in a state of humility of not knowing what the tradeoffs are and therefore what the thresholds should be.

Beena: In complete transparency, I used to work at GE, worked across their aviation, locomotive, power, and that was the primary use of what we did with data and AI was predicting a lot of these engine failures. I almost think that we put a lot of focus on bias and associating ethics and bias almost interchangeably, whereas ethics is so much more than depending on the use case or the industry. There is so much nuances on how that AI or the algorithm might be used depending or even whether it's healthcare, the AI you might be using to predict, say an MRI machine failure versus the AI that's being used for hospital bed management versus patient care, each one of those beyond the value creation the unintended consequences are going to be very different and then to your point setting that right threshold is where the design principle comes into play is which one is it more acceptable? Stuart, this is such a fascinating discussion. I come from a background where I have worked with banks and financial services, with trading and manufacturing, and I have just seen there is so many nuances around AI that we tend to try to bucket it all under one umbrella and I think it's time to your point, think about what is the real objective and it is going to be different depending on the algorithm you are using, but to what intent and to be really clear on that objective and the details around it.

Stuart: I think you have to be careful, not just to think about objectives, but to think about the whole context of deployment. One of the things that happens with algorithms is that they influence their own data environment. This is well known in credit and insurance that by offering loans with no requirement for any guarantees, you create the incentive for people who have no interest in paying back the loan to apply for the loan. So, you create a selection bias in insurance, it's called moral hazard. So the operation of the algorithm ends up changing the data that comes into the algorithm and then you have to think about where does that dynamical system end. Does it reach a new stable equilibrium that's good for us and good for the customer or do we go bankrupt or what? So thinking about this not in this sort of very isolated black box way. I have got my training data, I have trained my algorithm, I got 92% correct, so I am going to deploy it, that seems like a recipe for disaster. In fact, there are a number of recent papers, I think the most prominent came from Google, it's a paper about under-specification. They looked at 30 deep learning systems that had performed extremely well in training and testing and then when they deployed them, they failed miserably. This notion that you can simply test a system in isolation and see how good it is, and that's a meaningful number, it turns out not to be true. The EU in preparing their new round of regulations on AI, which they published in draft version, their earliest versions focused almost entirely on this idea that all AI systems should be submitted to national testing centers where they are going to be tested for accuracy and then sent back with a stamp of approval, and it was like what are you talking about, this could not possibly work.

Beena: So true. I think you touched on this piece, working in the real world. Again, it's not just about the accuracy of the algorithm, but in this new world where we are thinking about human compatible AI where humans play an active role, how do you train this workforce so that they are engaging the right way or are really leveraging the power of AI to benefit them. People like you and thought leaders like you obviously get that. But AI is out there in front of everybody, how do you bring the rest of the population or the rest of the workforce up to speed so that they are leveraging AI to truly benefit them. At the end of the day, it's not just people who are building and designing AI, it's also the end users who need to be using. Any thoughts on that Stuart?

Stuart: I think of this almost more from the user point of view. One canonical deployment situation is the domestic robot, which is still some distance away, we still don't have anything close to the perceptual capabilities, the manual dexterity, and the common sense, and the planning capabilities to actually have a domestic robot that could be really helpful let's say to an elderly person or a disabled person in enabling independent living. But obviously it's something that would be enormously valuable. Now that system when it comes out of the box doesn't know who you are and it can learn something about you from where you live and maybe even have a little conversation with you just to get to know you a bit, but it's going to have to learn a lot about your preferences for how things work and if we do this right, a human user shouldn't require any special training for doing this, it should be as sort of a natural interaction as we do with children for example. Sometimes we say, don't do that, do it this way. Sometimes we show them how we want them to do things in a positive way. Sometimes we are just happy with things that they do or disappointed, and all this kind of feedback it sort of works for children and I think it will it will work for humans that are working with robots or software AI systems. I like your view that it should really should really be about augmenting humans, it should really be think of it as putting power tools in the hands of every human, and if we do that right, everyone becomes more productive and therefore more valuable, more valued, and that's good. I am afraid that that's not the

way a lot of people think about it. A lot of people think about it in terms of replacing humans. Clearly there is a strong economic incentive to do that and some professions inevitably are going to shrink.

Beena: Stuart, this has been great, how can our audience stay connected with you, where can they find you, and follow your amazing work?

Stuart: That's very kind of you Beena. So the two books that I mentioned, one is called Human Compatible and that's a non-technical book that explains some of the ideas I talked about in terms of the new model for provably beneficial AI and then the technical book is the textbook that Peter Norvig and I wrote called Artificial Intelligence: A Modern Approach. That has a website which is ama.cs.berkeley.edu. You can find me, I don't do Twitter or Facebook unfortunately, but you can just Google me and then my Berkeley web page will come up and that has links to publications, as well as online talks and lots of newspaper articles and other things.

Beena: Stewart, thank you so much. We will include links to everything that you just mentioned in the episode notes so that our audience has access to it. I really enjoyed our conversation and learned so much. Thank you again for your time. Take care.

Stuart: It's a pleasure. Thank you Beena.

Beena: For our listeners, be sure to stay connected with AI Ignition at the Deloitte AI Institute. Thank you.

Thanks for tuning in to another episode. Check out our AI ignition page on the Deloitte AI Institute website for full video and podcast episodes, and tune in next time for more thought-provoking conversations with AI leaders around the world. This podcast is produced by Deloitte. The views and opinions expressed by podcast speakers and guests are solely their own and do not reflect the opinions of Deloitte. This podcast provides general information only and is not intended to constitute advice or services of any kind. For additional information about Deloitte, go to deloitte.com/about.