Transcript TrustTalk podcast interview Benjamin Kuipers

Voice-Over: Welcome to TrustTalk. Today's guest is Benjamin Kuipers. He is a Professor of Computer Science and Engineering at the University of Michigan. His research in artificial intelligence and robotics focuses on the representation, learning and use of foundational domains of knowledge, including knowledge of space, dynamical change, objects and actions. He is now investigating another important foundational domain: ethics, trust and cooperation for robots and other AIs that may act as members of human society. Your host today, Severin de Wit.

Podcast Host: Benjamin, welcome to TrustTalk. You worked for many years in artificial intelligence with a focus on the nature and the acquisition of common sense knowledge. Now, before we dive into that, I'm intrigued by the words "common sense knowledge". What does that mean?

Benjamin Kuipers: Thank you very much, Severin, for inviting me to be on TrustTalk and I appreciate that nice question. We each have some degree of common sense knowledge, for example, of navigational space because we use that to travel between home, work, friends, shopping, and places like that. Of course, you and I live thousands of miles apart in different environments, so the cognitive maps that are part of each of our common sense are describing different environments, different pieces of geography. So common sense is our ability to represent our own knowledge of space. It's not a shared knowledge about all spaces. And so this same principle applies to other what I call foundational domains, like thinking in terms of objects, actions, dynamical change, theory of mind, ethics, and so on. So by studying the common structure of this kind of foundational knowledge, we can learn how humans and robots can learn spatial knowledge from their own experience navigating through the world and learn how to plan routes to get to their desired destinations, home, work or whatever. And the structure turns out to include things like travel paths, decision points, connectivity of the route network, local and global frames of reference, and so forth. There's a lot of individual variation. Some people constantly keep themselves oriented with respect to the cardinal

directions (north, south, east, west). Others don't. Some people discover shortcuts, other peoples never do. These differences give us clues about how spatial, common sense knowledge is represented, learn, used and explained. Now it's easy to see how common sense knowledge of space is helpful to the individual. Now, when I started looking at ethics as a kind of common sense, it became clear that the benefit of ethics is really for the society the individual belongs to and the individual benefits when their society thrives. And of course, they do badly if their society doesn't thrive. I'd like to just point out that the things that I am commenting on are really drawing heavily on wonderful works by wonderful thinkers in philosophy, biology, psychology, sociology, economics, business, and a lot of other areas. And I use them as needed as an AI researcher.

Podcast Host: Intelligent robots, for example, autonomous vehicles and other artificial intelligence like, for example, high-speed trading systems, make their own decisions about the actions they take. And as a result, you and colleague computer scientists take the view that those robots can be considered as members of our society, right?

Benjamin Kuipers: Exactly. If an artificial agent has goals of its own and makes action decisions based on those goals and it has a model of the world, then it participates in our society. For a society to thrive and perhaps even for it to survive, its members need to contribute to its ability to defend itself against threats and to take advantage of opportunities. Now we're talking a lot about robots, but they're not the only artificial agents that participate in our society. Large corporations, for example, whether they're for-profit corporations or not. They're artificial. They have goals of their own and they make plans and act to achieve them.

Podcast Host: Humans need morality and ethics to get along constructively as members of our society. So if robots and creations of AI are human, it is logic we expect them to also behave ethically. And if we stick to the subject of this podcast, expect him to be trustworthy.

Benjamin Kuipers: I wouldn't go so far as to call robots and AI's humans. Terminologically, I reserve the word human for biological members of the species Homo sapiens and I used the word agent when I'm talking about the larger category that includes both natural and artificial agents. There is intriguing evidence of innate cooperative tendencies in human infants and toddlers. Innate knowledge was learned by the species over evolutionary time, so we can think

about it as a kind of learning and the individual gets it becomes available after birth. Still, even so, we expect that children will need to be instructed in how to behave towards others and how to be trustworthy. We should certainly expect the same thing for robots, AIs and corporations. They will certainly not spontaneously become trustworthy. But then neither do we.

Podcast Host: So let's dive a bit more into ethics. Your work concentrates on a computational view of the function of ethics in human society. And in your work, you discuss its application to three diverse examples. So for our audience, what do you mean by a computational view of ethics? And which examples can you give to show the role of ethics and trust?

Benjamin Kuipers: Well, thank you. This is going to take a little bit of telling, so take a deep breath, because I'm going to be following a causal chain backwards, starting from the success of the society that we share to cooperation, to trust, to trustworthiness, and to the role of ethics. So how does a society benefit from the activities of its individual members? Game theory provides mathematical tools for describing various interactions among individuals. This is a very rich and complicated area, but one of those tools classifies interactions as positive-sum, for example, win-win interactions, zero-sum (like most recreational games) and negative-sum. If I beat you over the head and take your wallet, then I get a small gain and you get a big loss. So when interactions are mostly positive-sum society comes out ahead in the long run, regardless of who wins that game, we can apply the term cooperation to the wide variety of positive-sum interactions. Think about farmers getting together for a barn raising or to bring in the harvest. A lot of people work together to produce a great value for somebody with the assumption that other people are going to chip in as well. Cooperation, this next part, requires trust among potential partners. If you're cooperating, you're vulnerable to your partners. They might fail to contribute as promised to the collective effort. They might take more than their share of the resulting gains. But what is trust? There are many definitions of trust, but the most relevant one for my purpose comes from business management. This is a restatement, trust is the willingness to accept vulnerability to another, confident that the vulnerability will not be exploited. By accepting this vulnerability, cooperative effort can pay off far more than the sum of individual efforts. So if I go and help somebody raise a barn, I trust that they're going to help me raise my barn sometime. I'm vulnerable to them failing to accomplish that. But if each of us is willing to do our part, we all do better than we would otherwise. So that example shows that trust is warranted when other people are trustworthy, or when other agents, other actors are

trustworthy. If the others are worthy of trust, then I can count on them to keep their promises about what they'll do. And I can count on them not to take advantage of an opportunity to seize more than their share of the fruits of our common effort. By counting on trustworthy others, my planning is greatly simplified. I don't need as many resources to defend myself against those others or for recovering in case they exploit me. Now, this is true, even if this cooperation isn't with a particular identifiable set of partners like, say, the other farmers in my community. For example, one kind of cooperation is driving on the correct side of the road or stopping when the traffic light turns red. Those are cooperative acts where the partners are all the other drivers. Each driver accepts some small sacrifice or vulnerability in return for the much greater benefits of safe and efficient transportation.

Podcast Host: So how do I know to be trustworthy myself and how to recognize trustworthiness in others?

Benjamin Kuipers: Well, in my view, this is the function of ethics. Ethics is a body of knowledge that a particular society has at a particular point in its history. And it's used to instruct the individual members of that society on how to be trustworthy and how to recognize trustworthiness in others. If we look at societies over history and geography, we see that the ethics of a society changes dramatically, but mostly over the centuries, but sometimes over faster time scales. If a society has widespread trust, trustworthiness and cooperation, it will naturally have more resources available for defending against threats, pursuing opportunities, and is more likely to thrive. But if you have a society whose ethics encourages exploiting the vulnerabilities of others, it's less likely to thrive.

Podcast Host: So like you explained, robots are no longer just made to perform tasks but are increasingly made for social interaction. How did we get here by thinking about ethics and trust for robots?

Benjamin Kuipers: That's a very good question. So thank you. Thinking about how to design a robot that includes the foundational domains of ethical knowledge, just like we thought about cognitive maps and finding our way from one place to another, we found ourselves asking about the purpose of ethics and its functional role in the life of an intelligent agent. And then this led us on a path from the individual to the society to positive-sum interactions to trust and

trustworthiness, and finally to the ethics taught by our society. Now, if you think about interacting with an untrustworthy member of our society, you can't count on them to fulfill their commitments, and you need to allocate extra resources to defend yourself and perhaps to recover from their misbehavior. If this is a robot, you're not going to buy it. So from a marketing perspective, trustworthiness in a robot is absolutely essential.

Podcast Host: You explained how ethics in AI, in fact, acts as a body of cultural knowledge that encourages individual behavior promoting the welfare of society, which in turn promotes the welfare of its individual members. Like you said, trust plays and cooperation plays a key role here. In many countries, however, there is a trend towards less cooperation and sharpening of differences. Have robots a role to play to bring back cooperation?

Benjamin Kuipers: Oh, yes, I think so. But let's back up a little bit. Think about the triumph of modern genetics. One foundation for modern genetics was the detailed study of the genetics of Drosophila, that is fruit flies. That happened not because people wanted to breed better fruit flies, but because fruit flies are methodologically convenient for doing experiments in genetics. So building useful, intelligent robots also is a convenient and necessary way to investigate and experiment with the roles that trustworthiness, trust and cooperation play in the success of our society. The understanding that we gain by trying to develop robots, applies not just to those robots, but also to humans and to the corporate entities that also participate in our society. Humanity, all of our societies face existential threats, including nuclear weapons, infectious diseases, and climate change. Meeting these threats will certainly require global cooperation, which, of course, is going to require global trust and trustworthiness. Now, as you point out in your question, there are strong ideological trends out there that encourage profiting by exploiting the vulnerabilities of others, individuals and groups. The profits from strategies like that, exploitative strategies come from negative-sum games. The loser loses more than the winner gains. This discourages trust and cooperation, and therefore it makes the society weaker and less able to meet those existential challenges.

Podcast Host: Psychologists have indicated that much of the mistrust of humans of robots is caused by what it is called the "Frankenstein Syndrome", the fear that a robot made by a human can rebel against its creator or derail its function and cause harm. Robots, especially humanoid ones, are also considered outsiders, which is why humans naturally tend to ostracize

them. As robots grow even more autonomous, the mistrust from humans seems here to stay. How do you see science can restore trust in robotics?

Benjamin Kuipers: Well, that's an interesting and long-term question. Think about the fact that we as humans try not to give children more responsibility than they're ready to take on. So, for example, there are legal age limits on driving and voting, and parents judge carefully how much money a child should have to spend, as they learn how to handle this kind of responsibility. Fictional scenarios that explore what happens with out-of-control robots often involve a robot given more power than it can handle responsibly. There are a lot of these fictional examples, but in one case in the "Terminator" series, the system Skynet is given control of the United States nuclear arsenal, and when its human designers try to unplug it, it triggers a thermonuclear war. In another movie, very different, called "Robot & Frank", a companion robot who is keeping a retired jewel thief company, encourages him to undertake another theft, basically in order to keep him psychologically engaged for his mental health. This does not end well, although it's a very entertaining movie. Now, in the real world, autonomous cars have crashed into stopped emergency vehicles that any driver would have seen and avoided. Disembodied A.I. systems sometimes cause "flash crashes" on the stock market, or they make fraud or bail or sentencing decisions based on overly simple models that do damage to people. These systems have been given way too much power in the real world to affect real people's lives before they have demonstrated their trustworthiness. So as robots and AI systems are increasingly deployed in our society, we need clear ways for them to earn our trust. And so the kind of research that a number of people are working on, including me, helped give us a growing understanding of the role of ethics, trust and cooperation that needs to be applied to corporations and to humans as well as to robots.

Podcast Host: Coincidentally, you just mentioned children. A scientist from the University of Amsterdam, Chiara de Jong, defends today her PhD thesis called "Children and Social Robots: Towards a Better Understanding of Their Acceptance of a New Technology". Most of the research on child-robot interaction focuses on the possible learning gains that come from the interaction. However, in order for an interaction to be successful, the child first needs to accept and trust a robot.

Benjamin Kuipers: That's fascinating work, and I really look forward to hearing about what she's accomplishing. Now, there are several other people I've encountered over the years. A psychology professor at the University of British Columbia named Kiley Hamlin showed, starting with her PhD thesis at Yale, that pre-verbal children have a strong preference for supportive or cooperative characters and a dislike for obstructive or exploitive characters. So this is a very early time for them to be perceiving moral value or unvalue in the environment. Similarly, psychology professor Felix Warneken, now at the University of Michigan, showed that toddlers will demonstrate a strong tendency to spontaneously provide cooperative help when they observe an adult who needs some kind of assistance, even when he's not making any request or giving or promising any kind of reward. So these and similar findings suggest that human children have a strong innate bias towards trust and cooperation with others. Allowing a child to observe a robot being helpful and cooperative toward others is likely to encourage that child to trust the robot.

Podcast Host: My last question, Benjamin. In what way will AI and robotics change our life, say, in ten years' time?

Benjamin Kuipers: I think that is a fascinating question. And I need to start back a while ago. In 1973, I entered graduate school in pure math, and that's when I first encountered artificial intelligence as a research area. I was overwhelmingly excited and I said to myself and my friends, we're going to solve the problem of the mind in five years, ten years max. As you undoubtedly noticed, it didn't happen that way. So over the decades since then, I've seen repeated intellectual revolutions in artificial intelligence. These provided exciting results and genuine progress, and everybody gets all excited about what's going to be happening very soon now. But then each one reaches a plateau and we start to realize that we need something more. This has led to a certain amount of discouragement from some people. But when I look at this, I think about the progress of modern physics. So modern physics really started with people like Newton and Leibniz about 350 years ago. And we're a long way from done. We made a lot of important progress, but we're not done. Modern work in artificial intelligence started about 70 years ago, and in my opinion, the problem that we're working on is at least as hard as the problem of physics. This is the problem of the mind and how the mind works. So in ten years, we're going to see exciting progress. Very likely we will see autonomous trucks driving cargo between terminals in the outskirts of major cities. But I would be skeptical that

you're going to see very many autonomous taxis driving into the neighborhoods of many cities. You may see it in some. But it will not be as widespread as people now think. Telephone answering systems, the little voice things that listen to what you ask for, they'll improve from where they are now. But they will still be very frustrating because they don't actually understand what you're asking. Ads that you receive will be better targeted than they are now, but you may still be worried about privacy and who is drawing inferences about your life and how much they know, and what they might do with that information. We will also still be in the middle of climate crises and political crises. And we hope that we will have learned better how to encourage trust and cooperation, especially among the humans and the corporate entities that will have such huge influence on the way our world develops. Robots and AIs will be there, but they'll be marginal, relative to the impact of human decision-making and corporate decision-making. Scientifically, I believe it's an enormously exciting time to be working in artificial intelligence. But we're working on a problem for the centuries with progress over decades. This is not a problem that gets solved in a decade or two with huge progress year by year.

Podcast Host: So with my question about ten years, I'm way too optimistic.

Benjamin Kuipers: Yep. Actually, let me encourage you, write down your predictions and then look at them.

Podcast Host: and see what happens in ten years, right?

Benjamin Kuipers: Let's see what happens in ten years.

Podcast Host: Well, Benjamin, you have a fascinating area of expertise and science, and thank you very much for your insights. It's especially for laymen, it's difficult stuff to understand, but it's also fascinating. And you have helped us understand a bit more on the subject. Thank you for that and wish you all the best and good luck with your science and future discoveries in AI.

Benjamin Kuipers: Well, thank you very much, as you can tell, I believe that trust is one of the most important concepts in this particular domain. And so I appreciate your giving it very careful attention.

Voice-Over: We hope you enjoyed this episode of TrustTalk. We would be very grateful if you would leave us a review on Apple Podcasts or on Stitcher. Don't miss out on Future Travels around trust and subscribe to this channel or visit us on our website TrustTalk.co or on Twitter at TrustTalkCo. We look forward to seeing you again.