**TIANMIN:** So, I think a lot about not only building artificial intelligence, but also how human intelligence can inform us to build better AI. And I think that the idea of building a world model that actually originally came from cognitive science and then specifically I have a strong interest in the social aspect of AI and also human intelligence.

**TIANMIN:** So, how we can build human level kind of social intelligence so that we can allow systems to understand humans and really cooperate with humans in the same way that we can understand and cooperate with each other. We have our thoughts in our minds, and we think about those thoughts in the form of language. We can also write down those thoughts and use, use our language to communicate with others about our thoughts.

**CRAIG:** Hi, I'm Craig Smith, and this is Eye on AI. In this episode, I speak with Tianmin Shu, a rising figure in the AI research community soon to join Johns Hopkins University with an impressive background that includes a postdoc at MIT and a PhD from UCLA. Tianmin works at the intersection of AI and cognitive science with a particular focus on societal aspects of both human and artificial intelligence. Our conversation revolved around the concept of world models, a term rooted in developmental psychology and cognitive science and their integration into AI. Tianmin elaborated on how these models coupled with large language models can enable the creation of agents capable of interpreting and interacting with the world in human-like ways. He sheds light on his projects involving world models for household environments and the potential of multimodal sensory data in their training. The episode also delved into the challenges and future trajectories in AI, especially in the realm of social learning, where AI models learn alongside of and from humans. Tianmin discusses research driving the future of AI and human robot interaction. And I hope you find our conversation as insightful as I did.

**CRAIG:** At home and work, we all know one person who is password challenged: sticky note reminders, emailing passwords, reusing passwords, using the word ‘password’ as their password. Because data breaches affect everyone, you need 1Password. 1Password combines industry leading security with award winning designed to bring private, secure and user friendly password management to everyone. Companies lose hours every day just from employees forgetting and resetting passwords. A single data breach costs millions of dollars. 1Password secures every sign in to save you time and money. 1Password lets you switch between iPhone, Android, Mac and PC with convenient features like autofill for quick signups. All you have to remember is the one strong account password that protects everything else - your logins, your credit cards, secure notes, or the office Wi Fi password. 1Password generates as many strong unique passwords as you need and securely stores them in an encrypted vault that only you have access to. I use 1Password and you should too. 1Password’s award winning password manager is trusted by millions of users and over 100,000 businesses from IBM to Slack. It beat out 40 other options to become Wirecutters top pick for password managers. Plus, regular third party audits and the industry's largest bug bounty keep 1Password at the forefront of security. Right now, my listeners get a two week free trial at 1Password.com/eyeonai.

**CRAIG:** So Why don't you start, Tianmin, by introducing yourself and then I'll ask questions from there. Give a little of your educational background. Yeah. And where you're working today and what you're working on.

**TIANMIN:** So I'm actually going to join Hopkins the CS department and also the Cognitive Science department as an assistant professor.

**TIANMIN:** But previously I did my postdoc at MIT and before that I did my PhD at UCLA. I' working at the intersection between AI and cognitive science. I think a lot about not only building artificial intelligence, but also how human intelligence can inform us to build better AI. And I think that the idea of building a world model that actually originally came from cognitive science and then specifically I have a strong interest in the social aspect of AI and also human intelligence.

**TIANMIN:** How we can build human level kind of social intelligence so that we can allow systems to understand humans and really cooperate with humans in the same way that we can understand and cooperate with each other. And again, I think that's something that's is very challenging right now.

**TIANMIN:** And also it's not something that big companies are working on. And I think it's also something that will really benefit from having a world model. Both in the sense of understanding how humans understand the world and also how we can understand humans that live in this world.

**CRAIG:** Okay. Can we start first of all, by talking about world models?

**CRAIG:** I've had a few episodes on world models. But maybe you can, when did the idea of world models emerge? I've been following Yann LeCun's work, but I know it's much broader than that. When did that idea emerge and can you give us just a very brief history of its development?

**TIANMIN:** Just a to determine, I don't think the history, I can tell. Probably not the full history. There's going to be some people at some point talking about world models that I don't know about. But the kind of case I'm familiar with is coming from developmental psychology and also cognitive science.

**TIANMIN:** So in developmental psychology there's this theory about core knowledge. So it's actually a theory. Proposed by Liz Spelke, who is now a professor at Harvard. So she had this theory about so since first, we have some core knowlege about the world and about other agents. that we can use as a foundation for developing more sophisticated intelligence that we can develop later in life.

**TIANMIN:** And one of them is to understand the physical world. For example, we have some physical common sense like object permanence, we understand supports if you put something on a table, it won't fall. So this kind of basic kind of physical common sense allows us to actually develop more sophisticated intelligence.

**TIANMIN:** And we have acquired those critical common sense very long, maybe even through evolution. Now it's unclear that how much of it actually comes from evolution. How much of it comes from experiences living life. Now there's also this idea about having an internal world model that can allow you to simulate what could happen in the future.

**TIANMIN:** And you can use that kind of simulation to not only reason about future physical states, but also applying yourself better through this kind of simulation. And that comes from a lot of work from my PhD, sorry, not PhD advisor postdoc advisor, Josh Tenenbaum. He had this idea that human has this internal intuitive physics engine. So we all know physics engine, which allows you to simulate what could happen in the physical world.

**TIANMIN:** Now human, in our minds, we can have an intuitive physics engine that allows you to simulate what could happen in the physical world, just like how you can simulate what could happen in a physical world using a game engine. Now using that, they show that you can simulate physical events.

**TIANMIN:** And use that kind of simulation results, adding some kind of noise on top of that. And you can use that to approximate how human judge physical events pretty well. They did experiments where they show a tower of blocks, and they ask people to judge whether the block will fall or not.

**TIANMIN:** If they will fall which way they will fall to, right? And then it turns out that if you actually simulate that using a game engine, you And as a little bit of noise on top of that, because people have imperfect imperfect kind of simulation, right? Physical simulation plus noise or some kind of uncertainty, that's approximately human judgment really well.

**TIANMIN:** And I think there's also recent kind of collaboration with a neuroscience at MIT Nancy Kanwisher. They show that there's some kind of evidence in our brain that maybe our brain is also using actually a game engine to simulate what could happen in in a physical world.

**CRAIG:** Yeah.

**CRAIG:** And that engine to, to simulate what's the computer code or algorithmic architecture behind that?

**TIANMIN:** Yeah. So if you if you want to write computer code there's there's this tools called probabilistic programs. I don't know if you have heard about it. Probabilistic programs.

**TIANMIN:** I think one way to understand it is that you can generate you can write down codes to describe a generative process. In this case, you can describe a generative process as a world model. You assemble objects, you simulate force. And then the generative model will tell you, after you apply this force, what will happen next. What will be the state of these objects, you just simulate.

**TIANMIN:** You can write down this generative process as a program. And then when you run this program, you will sample this sample foundation model. And it will tell you what's the distribution of the, for example, in this case, what's the distribution of the future states of these objects after you run the simulation.

**TIANMIN:** So that allows you to, first of all, simulate the generative model of the world. And second, it gives you a distribution. And what you can do with this distribution is you can conduct probability inference. So say if you want to predict where these objects will fall in the next few steps you can use that simulation, that distribution to tell you what is the likelihood of these objects falling into this place.

**TIANMIN:** You can also use this simulation to do planning, for example, with uncertainty. I think probably this object is going to fall this place, but maybe also in that place. Based on my simulation, based on the distribution coming from the, this policy program now given this uncertainty, how I can plan myself in order to maybe catch these objects.

**CRAIG:** I see. And that, that is, that is the world model?

**TIANMIN:** Yes. So in that that's like one, one implementation, right? You use a physical simulation and that allows you to sample future states. And that can be implemented as a publish program.

**CRAIG:** I see. And then in your talk you were talking about combining world models and large language models and the language model would provide the reasoning and the world model would provide the the. predictions of future states, and then the language model would pick which action to arrive at a future state. Is that right?

**TIANMIN:** We explored or discussed different ways in which people have tried to use language model to implement as like a word model. So the most direct use is say, let's prompt the language model with the current state, the action, and ask the model what could happen next.

**TIANMIN:** So this is the basic idea of a word model, right? You want to predict what will happen given the current state and given your action. And then I think that the paper you mentioned is the reasoning as planning paper the RaP paper. So the idea behind that is that why not we try to use a language model to do this.

**TIANMIN:** To imagine what could happen after taking an action at the current step. It works to a certain extent. I think in some simple scenarios, it can work. I don't think at this point, the model is robust enough to serve as a model on its own for more complicated scenarios.

**CRAIG:** Okay, and then this most recent paper and the the workshop at NeurIPS was about combining the language model with a world model.

**CRAIG:** So the world model predicts the future states and the language model provides the reasoning. Did I understand that correctly?

**TIANMIN:** I don't think that's I guess the complete idea. The idea is that we want to have expressible models and also express agent model that's built upon our model.

**TIANMIN:** And then using agent model and world model we can conduct model based reasoning. Now we want to explore the idea that, how language model could be used as a back end to implement a world model and an agent model. And of course like the RaP paper, there has been effort that try to use language model directly as, for example, word model and agent models.

**TIANMIN:** That can work in certain scenarios, but not in more complex scenarios. Now we want to explore, for example, how it can maybe enhance the agent model to serve as a better world models and agent models. And second, we want to also explore ways in which we can not only just use language, but also use other modalities, like the vision even like touch or audio to to really build multi modal world models.

**TIANMIN:** That's, so obviously that's not going to be the kind of typical language model you see nowadays.

**CRAIG:** Yeah. So can you talk a little bit about that? How you introduce the language model into this into this architecture with a world model in order to produce an agent?

**TIANMIN:** So I think I would say there are probably like two ways you can introduce this one way.

**TIANMIN:** That's that's not coming from my own work. That's from My, my colleagues at MIT they introduced the idea that use language as interface to translate language discussion about scenarios into a probabilistic program so that you can conduct probabilistic inference using that probabilistic program.

**TIANMIN:** And again, probabilistic programs serve as a training model of the world. And also about other agents. This idea comes from very classical idea in cognitive science called language of thought. We have our thoughts in our minds, and we think about those thoughts in the form of language, right?

**TIANMIN:** We can also write down those thoughts and use, use our language to communicate with others about our thoughts. Then if you do the reverse process, we are given language about scenarios. And also a question about that scenarios, how you can conduct reasoning using internal thoughts, not as directly as a language reasongin, but you want to convert language into an internal thought and reason over that thought.

**TIANMIN:** For example, you can describe a language, you can use language to describe a physical situation. I just told you yes. Maybe a tower of blocks on the table. On the bottom you have a small block. In the middle I have a larger block. On top, you have an even larger block. I can ask you, can you imagine how stable this tower of blocks can be, right?

**TIANMIN:** And I also describe a different scenario. On the bottom you have a very large block. On the middle, you have a smaller kind of on top, you have a teeny tiny block. Then you can also imagine what could happen next. Notice I have shown you a zero image about this, but you can based on my language execution, try to understand what I'm trying to describe here, right?

**TIANMIN:** So how we can convert that language into something like a thought process that allows us to simulate what could happen next. So I think what one so the colleagues at MIT, they come up with this idea that you can use a language model to translate this standard distribution into code. And so basically a probabilistic program that will basically become a physical simulation.

**TIANMIN:** So the scenario I told you, after the model, you'll get a code that allow to simulate the physical situation, and you can simulate what could happen next. And that this again, allow you to do model based kind of reasoning about the physical world. So I think that's very neat idea. And it's also very good use of language model, because again, I don't think language model itself has all the knowledge about the physical world, but language model is really good at writing code.

**TIANMIN:** And we all know that. But, and again, you can translate do this inverse process where you can convert language into a thought.

**CRAIG:** Yeah that's really interesting. And then how, and you guys have done this isn't theory. You've built these models to work together.

**CRAIG:** What is the agency layer. That's the reasoning and the planning or decision making. Then how do you add the the agent on the model?

**TIANMIN:** Yeah so agent is model as a decision making process, basically in this framework where you have your your world model, your internal world model that you can use to simulate what could happen.

**TIANMIN:** You have your own goal and goal has a condition of the goal is you have a reward. So you really will try to help you to reach the goal, but also in the process, minimize the cost. And then you have your belief about the world because agents only has, an agent only has partial observation of the physical environments.

**TIANMIN:** You don't know everything. So you have a belief about what the actual physical state is. So again, belief is like distribution of the physical states. Now, based on this, then you can try to make a plan that can based on your, first of all, your belief about how the actual physical states look like and then based, and then also try to maximize your rewards that allow you to reach the goal after doing simulations about different kinds of plans.

**CRAIG:** Yeah, yeah. So I'm trying to think of it in. So that's an RL, a reinforcement learning engine in effect.

**TIANMIN:** That doesn't have to be reinforcement learning, but that can be a model based planning. So the idea about so the idea behind the model based reinforcement learning or model based planning is that you don't need to do a lot of trial and error in actual physical environment.

**TIANMIN:** Okay. You can do so using a simulation. You can simulate beforehand what could happen after taking different kind of actions. Then you can search for the best set of actions that allow you to reach the goal without actually trying out in a real physical world.

**CRAIG:** At home and work, we all know one person who is password challenged: sticky note reminders, emailing passwords, reusing passwords, using the word ‘password’ as their password. Because data breaches affect everyone, you need 1Password. 1Password combines industry leading security with award winning designed to bring private, secure and user friendly password management to everyone. Companies lose hours every day just from employees forgetting and resetting passwords. A single data breach costs millions of dollars. 1Password secures every sign in to save you time and money. 1Password lets you switch between iPhone, Android, Mac and PC with convenient features like autofill for quick signups. All you have to remember is the one strong account password that protects everything else - your logins, your credit cards, secure notes, or the office Wi Fi password. 1Password generates as many strong unique passwords as you need and securely stores them in an encrypted vault that only you have access to. I use 1Password and you should too. 1Password’s award winning password manager is trusted by millions of users and over 100,000 businesses from IBM to Slack. It beat out 40 other options to become Wirecutters top pick for password managers. Plus, regular third party audits and the industry's largest bug bounty keep 1Password at the forefront of security. Right now, my listeners get a two week free trial at 1Password.com/eyeonai.

**CRAIG:** That last part sounds like a little bit like AlphaGo where you you simulate various courses of action, and then you, you search for the optimal solution and make that solution? Is it similar in in architecture to AlphaGo? The AlphaGo it's not using a world model.

**TIANMIN:** Okay. So what is the core component of AlphaGo? It's the Monte Carlo Tree Search, right? So that's a planning algorithm.

**TIANMIN:** That, that is not RL, that is planning. And then, and the reason why you can do planning in that case is that you know the rule of the game already, right? So you know exactly this will be the next stage of the game after taking certain move, right? And that allow you to do forward simulation. You can think ahead for many steps and get better plans by doing this kind of imagination.

**TIANMIN:** Now. You mentioned that there doesn't seem to be a world model. There is a world model. The world model is very simple. It's just the rules of the game. And then you know exactly what could happen next. Now the more, and the, and I think MCTS for planning, obviously it's very intriguing because it's very general.

**TIANMIN:** In a sense that if you have a reward, you will have a model that allow you to do forward simulation. In theory, you can do anything. If you do enough simulation, you can get pretty good planning. And if you add a little bit learning, like AlphaGo, You can also make the search even faster. Now the down, now the downside, the challenge is that in, in the real world, you don't have that perfect word model.

**TIANMIN:** But if you do build a world model for certain domains, then you can plug in that world model to do Monte Carlo Tree Search or any kind of model based planning.

**CRAIG:** Yeah. How far have your implementations of this gone? You've you've built a model with all these elements. Does it depend on Does its efficacy depend on how large the model is as is the case with language models, that the larger the model, the better performance, or is this more a matter of of just optimizing hyper parameters and stuff like that?

**TIANMIN:** You mean like building, actually learning a world model that can be used for learning a world model that can be used for very broad set of domains and not just specific domains. I think we are not remotely there yet. The reason is, for example there's just so many different kind of thing you can model in the world, right?

**TIANMIN:** You have so many different objects and even if you think about one specific object, they have so many different kind of properties you can try to model like the materials the textures the weight of the objects. The shape and all these different kind of properties, right?

**TIANMIN:** And depending on your task, your mind needs to model the object in different ways and to do this you need to find replications of the objects. For instance, if you only want to say carry some box around, you'll probably only need to know the weight and the rough shape of the box. You don't need to know very detailed kind of representation of the box.

**TIANMIN:** If you want to manipulate fluids, for example, I don't think shape and weight will be enough. You need to actually model fluid as some kind of particles. And how the particles or what flow interact with different kind of, with different kind of surfaces, right? If you want to say how I can manipulate some kind of deformable objects, so say if you want to cut some vegetables then you also need to know like how the vegetables will interact with the the knife.

**TIANMIN:** And then how they will maybe in some case, maybe a roll on the cutting board. So you need to somehow hold it. So there's a different kind of representation for this. If you only think about single objects. Now, if you think about wider case where you have many different objects in the world that's not only having different kind of properties, but also being also they are dynamic.

**TIANMIN:** So the states are also changed by the forces in nature, but also other people. So for example, objects may be moved by another agent while you're not present. So now you need to also imagine where the object could be. So world model will become very complex in the reward. Now how we can build that?

**TIANMIN:** I don't think there's a single answer yet. We have world models that can work pretty well in certain domains. So I mentioned that for robot manipulation, for example, we can, learn to how, learn how some kind of substance or methods, objects will change how their state will change given certain actions of the robots.

**TIANMIN:** And using that, you can try to manipulate robots for certain tasks. I think you mentioned Gaia. So Gaia is another great example of a domain specific kind of world model. Of course you can use that to maybe build self driving cars, but you cannot use that for, say, objects manipulation for robots.

**TIANMIN:** We have other cases where, for example, you can learn a world model for Some kind of abstract kind of knowledge. So there's some world model that's represented as a knowledge graph. So say if I do this what will happen, say, if I want to take a flight, I need to buy a ticket first. So some kind of abstract kind of common sense knowledge. Now you can use that to do certain things.

**TIANMIN:** So say like maybe help you to make a plan for a vacation. But again, that's because it's representing it as like an abstract knowledge. You cannot use that for physical manipulation. So I would say, I don't have a great answer about say, building a single world model that can represent anything in the world.

**TIANMIN:** I also don't think At this moment it's the best strategy to try to build such a model. It's probably better to say a world model that can do well in a range of specific domains. That's, and also you can easily generalize that to similar domains and fine tune it with a small set of data. But I don't think it's the best strategy yet to build one single word model that can do everything.

**CRAIG:** And the models that you've built are are trained or tuned for what domains?

**TIANMIN:** So specifically we've tried to build word models for household domains, so we have built at MIT, in our group at MIT, we built a household simulation called Visual Home. So we can build different kinds of apartments.

**TIANMIN:** You have different kinds of objects in an apartment you can interact with. You can put many different kinds of agents, simulated agents in those environments that again, you can interact with. So in those simulations, we can generate embodied experiences as training data to fine tune this for example, language model or different or other type of model as a world model.

**TIANMIN:** So this is a very different from say training a language model on text. And hoping you can get some knowledge about the words from those tasks. Because, again just intuitively speaking, for humans to understand the words, I don't think you can just read books and you can perfectly understand what could happen.

**TIANMIN:** There, there are many details about the world that won't be described in text. Say what happens if I push push this button of the dishwasher? Or what happens if I open the cabinet door, put another object into the cabinet and then go away a couple of days later, come back, how many objects will be there in the cabinet.

**TIANMIN:** Those kind of knowledge are very common to us, but you won't read it from books. But again, those are common sense knowledge that even kids can have. But if you but the idea here is that if you build embodied agents in those simulation environments, ask them to do different kind of tasks, interact with different agents, or even just do random exploration, you'll get these experiences, right?

**TIANMIN:** You'll get say what will be So how many objects will be in the cabinet after I did this sequence of actions? What will happen after I push this button of the dishwasher? What happen if I say, we haven't done that yet, but let's say what happen if I talk to someone and then I tell them that I'm pretty happy today let's hang out maybe in the afternoon will be the response.

**TIANMIN:** But if you can simulate all this and get a lot of this experience data, then we can train a model that can really understand how the world works or even how people will react to actions.

**CRAIG:** Yeah. And for training in a specific domain, is that done through video or still images or text or what is the data that you're using to train, for example, a household.

**TIANMIN:** Technically speaking, you can use any kind of data that you can get from the environment. So we have the vision data for sure. You can have the ground truth state of the objects. And agents, you can translate those ground truth states into language. So say, how many objects are there inside these cabinets?

**TIANMIN:** You can also there's some simulation that allow you to also simulate audio. So you can say if I drop this ball onto the ground what will be the sounds that I can get you can even simulate touch in some systems. I think in the industry, actually in the same group at MIT, they had the collaboration with, another group that works on multimodal sensory. So they have this nature paper in recent years where they build this glove that can allow you to simulate allow you to get touch sensory. So now imagine if you have some kind of VR set up where people can using this, using the kind of graph interact with different kind of objects and can give you the kind of even the touch sensors. Now you can use those different kind of multi modal sensory data to train your world models. We haven't gone that far yet, I think. So the work I tell you about we are using the ground truth state of the object and agents and then translate that into language.

**TIANMIN:** So we can more easily to train that, to understand the world. We have also done another another part that we have done is, so again, we are using states that we can extract from the images. Particularly to represent a state in the environment as a scene graph. So scene graph is the structure representation about the world.

**TIANMIN:** Each node is an object each edge connecting two nodes representing the spatial relationship between two objects. So if you have a cabinet, you have cabinet nodes. If a cup, you have a cup nodes. And then if the inside cabinet, there will be a inside the edge connecting the cup and the cabinet. So this kind of representation can be extracted from images but it's a pretty appealing kind of representation compared to pixels because from pixel you don't get this semantic knowledge about the physical states.

**TIANMIN:** Now you can, you don't need to represent it into language, but you can just use the symbolic representations of this scene graph as the world state, and again, you can train that model on top of those symbolic representations to predict what will happen next.

**CRAIG:** Yeah. During that training process I spoke to somebody a year or so ago, who was working on a training system that, that was hyper speed. So a hundred, I think it was a million frames a second or some credible rate. Do you have to train in real time or can you train at hyper speed? How long does a system need to absorb an image or frame a video?

**TIANMIN:** Yeah, so I think that depends on how data hungry your model is.

**TIANMIN:** And so say if you want to train model free RL, I think you mentioned this, right? So usually for training model free RL, you want your system to be as fast as possible because it is very data hungry. You have to try many different actions. just blindly, only just blindly, so that you can get enough data to, or enough training signal to update your policy.

**TIANMIN:** If you haven't tried enough, you probably have, probably cannot even get any positive reward at all in some very complex tasks. However, I think for training world models, on the other hand I think the speed is actually not the biggest issue. You can just, that's an algorithm, not algorithm even, just like any kind of embodied agent in the simulation.

**TIANMIN:** That is to explore, try different things in the environment for first a period of time. You're going to get a lot of data already. Because again, it's different from tuning our agents because our agent has the rule function that's defined for a specific goal. You try a lot of things, you get a lot observation data, but now those probably at the beginning are related to a task.

**TIANMIN:** So you don't get any positive reward at all. However for training world model you train on random things in the world, a lot of random thing will happen. And those random things happening is actually all the useful training signal for your world model, right? If I, if I. If I push something, something will be moved.

**TIANMIN:** If I push something down, something will be in a different location. If I walk from one room to another room, I know my location will be changed. My observation also will change. So that gives you many training signals you can use. So you don't need to, I think, compared to at least model free RL agents, you don't need to try too many steps to get useful data.

**CRAIG:** Yeah. And so you've trained in a household environment for that domain is this all in simulation or do you have you or is this outside of your area of research? Have you tried controlling a robot with this kind of a world agent model?

**TIANMIN:** I guess part of my research is also in human robot interactions.

**TIANMIN:** So down the line, we do want to evaluate all these models in the real world. Recently, I haven't done that yet, but there's there are other groups that have done that. So there's, I think we actually mentioned this paper in our tutorial. So there's a very recent paper from AI2, AI2 is the [Allen Institute for AI] AI Research Institute in Seattle.

**TIANMIN:** So they, they show that if you want to train a policy, a robot policy for indoor environment kind of navigation, so say, try to find a TV remote for me or try to train some policy to do simple, very simple object manipulation, say, pick up one glass and give it to me. Now and they train this actually on raw pixels.

**TIANMIN:** So they don't actually turn on ground truth world state. So give them pixel that we can see in front of you and then give the command, say find this object for me. You want the robot to actually navigate and search through the apartment and find the object for you. Now, previously people think Yes, could be a very big domain gap between simulation and the real world.

**TIANMIN:** In fact, I think many robotics researchers are actually against using simulation precisely because they think no matter how good simulation is, they're not going to be real. They're going to be a hundred percent real. And particularly if you think about the kind of policy we want to build here, it's crazy, right?

**TIANMIN:** It's actually low pixel inputs and they are then mapped to some actions. Now the pixel rendering from simulation will be definitely not going to be 100 percent real compared to real world images. However, they show that if you have enough data, you actually do not need any kind of real world data to fine tune the model you're training in simulation.

**TIANMIN:** And they can deploy in real world apartments, and that's the real physical robot to do the same kind of test. So I think that's very promising because the quality of the current simulation is already very good. That, that's one. And second, it allows to generate a lot of data, a lot of data you cannot get from the real world.

**TIANMIN:** And if you have enough data you can get very good policies out of it. And I think it also presents maybe even the third kind of opportunities. So when you have a lot of simulations you have household simulations, you have traffic simulations, you have those kind of objects manipulation simulation for robots, you have fluid simulation, you have other kind of physical fixed engine kind of simulation.

**TIANMIN:** You have all these simulation people have built. You have a model that can understand learn like general knowledge about the world from all these different simulations, then those that have knowledge are very useful for the real world. Even though like for individual simulation, they don't represent the whole world, but if you can combine knowledge from different kinds of simulation together, then you can have a, I guess like a forward picture about how the physical world works.

**CRAIG:** Yeah. And the this is all still in, in the research phase but how long do you think before, this is one kind of agent. There are a lot of different strategies to build agents, but how long do you, how promising do you think this method is as opposed to others? And how long do you think before somebody puts, refines this and puts it into a product in the real world?

**TIANMIN:** I guess like you mentioned Gaia, so I don't know exactly how they are using their simulation for their products. But they are startup companies. I don't think they're going to build anything that's just purely for research that have no real value at all. And then I think also we have things again I mentioned these indoor kind of robot navigation kind of tasks, right?

**TIANMIN:** Those are very useful right? If you want to build robots systems that can help you to do household tasks the, one of the most fundamental tasks you're going to ask a robot to do is to find objects you're going to need for whatever the thing you want to do in the house. Now down the line, if you train robots to do more complex tasks using similar kind of approach, I think will be a pretty compelling kind of product for you.

**TIANMIN:** And then that's definitely going to be more useful than Roomba. And I think that announcing that could be very useful for is not just embodied agents, like robots or self driving cars, but also just say, web agents. So think about works interface or any kind of software interface.

**TIANMIN:** As also a world model, like you can try to figure out how does the interface work after you've taken some actions and you can model it as a kind of virtual world model so to speak. Then if you can do that, you can also build any kind of system agents for web interfaces for softwares Et cetera.

**CRAIG:** Yeah. And so do you think that within a year or two years or five years there'll be whether or not they're embodied or virtual? There'll be, people will be working with agents built on this architecture in various domains.

**TIANMIN:** I am not good at predictions, but I think, embodied agents are always very hard to build. Like for example, you would think one of the most fundamental, most basic kind of task for robots, which is grasping an object. I think some robotics actually ?? says that whenever you start to work on grasping, you can have a whole career ahead of you.

**TIANMIN:** It's just a really hard problem right now. However, if you think about simplified tasks that doesn't require a lot of very compact physical manipulation say, maybe like web interfaces or some kind of virtual assistants that can help you to do some tasks. I think that could happen easily.

**TIANMIN:** I think there are already people building, trying to build such for that as well. And I think also maybe even say, for embodied agents. For very not very, but somewhat rigid environments. So say in I don't know warehouses or factories again, they are already products like that, but I think with better AI models, you can have more powerful kind of robot coworkers in those more rigid environments.

**CRAIG:** Yeah. But do you know of any commercial enterprise that's taking your research and implementing it?

**TIANMIN:** Right now, I don't think so. I don't have that knowledge. Yeah.

**CRAIG:** Yeah. It's fascinating research. So I see you refer to it as LAW,

**TIANMIN:** language model, agent model, and world model.

**CRAIG:** Yeah. I see that LAW. I don't know if that's a term you use. Yeah.

**CRAIG:** Okay. So what's next in your research?

**TIANMIN:** So like I said, building one models and also agent models. I think that's very fascinating direction for me. And also for many people now, and hopefully what people are finding fascinating and work on that.

**TIANMIN:** But I think another thing, and I'm really interested in is social learning, which was also talked a little bit about in the tutorial. So you have a model that can learn on its own, but think about humans actually learn a lot of things. with each other or from other people, right? And we can also teach knowledge we learn to other people.

**TIANMIN:** So how we can actually build models that can, or agents, that can also learn with humans, learn for humans, or learn from humans. I think that's another kind of direction I really want to work on.