Rationally Speaking #231: Helen Toner on, “Misconceptions about China and artificial intelligence”

Julia: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host, Julia Galef, and our guest today is my friend, Helen Toner.

Helen is the director of strategy at a new think tank at Georgetown University called The Center for Security and Emerging Technology, or CSET, which researches and advises policymakers on the security impacts of technologies like artificial intelligence.

Before that, Helen was a senior research analyst at the Open Philanthropy Project. She just got back from living in China for nine months, getting to know the AI ecosystem there. We're going to talk about a bunch of things, including common misconceptions about China and AI strategy, and just how to think about a topic like this, how to analyze complicated strategic decisions with a lot of uncertainty in them.

So, Helen, welcome to Rationally Speaking.

Helen: Thanks so much.

Julia: So, first, give us a little more detail on what you were doing in China -- aside from the most important thing you did there, which was, at my request, paying a visit to Beijing's replica of the coffee shop from Friends.

Helen: Mm-hmm.

Julia: I really appreciate you making that pilgrimage on my behalf.

Helen: It was a pleasure, Julia. I feel like that visit in itself says a lot about China today. I think, as I showed you in pictures, the coffee shop is a beautiful replica. They serve only foods that were mentioned on the show. They have the couch. Everything is set up beautifully.

Julia: And the walls, too. All the things on the walls are a perfect match for the show. It's a strenuous attention to detail.

Helen: Right, and it's on the sixth floor of, when I was there, basically an abandoned shopping mall.

Julia: God, that's so surreal.

Helen: In the indoors. So it was sort of in one way beautiful and perfect -- and in another way, it's sort of completely wrong.

Julia: Is that your encapsulation of – what, about China?

Helen: I mean, I think it's a little bit of an unfair encapsulation of China. I think certainly, there are many more shiny high-rises where it was a little bit unclear what a shiny high-rise was doing in that place than you'd find in the West.

Julia: So, “incongruity” really is the key word.

Helen: Yeah, and also this sort of strange techno-modern central planning, was sort of visible in various places, so this is an example. It's a shiny, modern, skyscraper but seems like someone just decided it should be there rather than there being a really strong business case for many businesses wanting to put a lot of money into this.

Julia: Right, right. I think this was you but it's possible it's another one of my friends who was in China recently: They were describing a similar experience of a sort of perfect replica of something in the West but with some of the details slightly off. I think it was a Starbucks --

Helen: Yeah, yeah. So, this was me. It was in Shanghai. I'd been living in Beijing, which is a little bit less foreigner-friendly, and Shanghai has more Westerners and so there's more Western stuff. They have the French Concession, which was previously under French control, so it has nice pastries and things.

So, it was a lovely coffee shop with a sort of high ceiling, this polished wood, great espresso machine. You were able to choose

where your milk came from, which I've actually never encountered anywhere else. Anyway, a very hipster coffee shop --

Julia: And as an Australian who had been living in the US, I'm sure you were missing your well-made coffee.

Helen: Absolutely. And they were playing, if I remember right, the Star Wars soundtrack, which was just incredible.

Julia: That's right. Wow! Exactly. Incongruity. That's what it is.

But anyway, I'd sidetracked us dreadfully there. My original question was going to be not what weird coffee shops you'd gone to, but what were you doing day to day in China? What was your reason for being there?

Helen: Yeah, so my day job, as I thought of it, was being in an intensive Mandarin Chinese language program. So, this was based at Tsinghua University, one of the sort of biggest and most respected universities in China. It's in Beijing. That was 20 hours a week, all small group classes.

Which I then sort of also supplemented, especially early on, for the first couple of months, I did a lot of self-study, as well. Which was super interesting. I really love learning languages. I could chat for a whole hour just about learning Chinese and what a great language it is and how fascinating that was.

So, as I said, I thought of that as my day job. So, then my side hustle was trying to meet with people in China who were involved in AI and machine learning in some way.

Julia: Oh, and I should just clarify for background that one of your main focuses, when you were a senior research analyst at the Open Philanthropy Project, was advising -- I guess, grantmakers or policymakers or researchers, about AI development. So, this wasn't coming completely out of the blue.

Helen: That's right. So, I'd been getting interested in sort of the policy side of AI at the Open Philanthropy Project. And specifically,

because I think “AI policy” is sort of this very, very broad term, specifically the national security angle, of what are the implications of this progress in machine learning that we're seeing?

China just immediately pops out in that field. If you start asking, "How should the US be thinking about AI from a security perspective," China is sort of the first word on everyone's lips here in DC. It kind of kept coming up in my conversations.

So, a big part of that trip -- part of it was personal interest, and just sort of being at a good moment in my personal and professional life to go spend a year overseas. But part of it was very much this sort of professional relevance.

So, that was what I was trying to get at, in my side hustle, which included setting up meetings with machine learning professors, and talking to them about how they were thinking about the longer-term future of this technology and the sort of social implications it might have.

Also, included, it's very easy to meet up with other foreigners, of course, and lots of people in China are really interested in thinking about AI, and sort of big ideas and the future of technology, and the future of the geopolitical order and so on. So talking to those people was also interesting.

I also had a fun time with my two language buddies who I found there. So, these two young women who I met once a week each and had lunch with, and would speak half an hour of English, half an hour of Chinese. They were both machine learning masters’ students. So, that was a really good way to just kind of talk to them about what their lives look like, how they thought about their careers and sort of all of that in a pretty low key setting. They became good friends, and that was a great opportunity as well.

Julia: That's great. So, from your experience both talking to AI scientists and also talking to young people in the field and just getting to know how AI in China works, is there anything that

surprised you? Compared to your preconceptions, or compared to media portrayals of AI in China?

Helen: Yeah. I mean, I think I've seen you write about this before, Julia, that sometimes people ask you about big ways you change your mind and, in fact, there are just lots of little adjustments. So, I think there are plenty of little adjustments and plenty of cases where I had sort of a vague view and it became much more concrete or much more detailed.

An example of a very sort of crisp change in my thinking is that I went in -- I feel like the West has this portrayal of the company Baidu as like the Google of China. When you think "Google," you think super high tech, enormous, has many products that are really cool and really widely used. Like Google has search, obviously, but it also has Gmail. It also has Google Maps. It has a whole bunch of other things, Android.

So, I feel like this term, "the Google of China" gets applied to Baidu in all kinds of ways. In fact, it's sort of true that Baidu is the main search engine in China, because Google withdrew from China.

But kind of all the other associations we have with Google don't fit super well into Baidu. Maybe other than it is one of China's largest tech companies. That is true. But just in terms of my overall level of how impressed I was with Baidu as a company, or how much I expected them to do cool stuff in the future, went down by a lot. Just based on, there's no ... Baidu maps exists but no one really uses it. The most commonly used maps app is a totally different company. There's no Baidu mail. There's no Baidu docs. There's a lot of stories of sort of management dysfunction or sort of feudalism internally. So, that was one of the kind of clearest updates I made.

Julia: Interesting. Why do you think that wasn't captured by the media? I feel like, Baidu's very high profile, journalists should in theory be uncovering stuff like that. Am I wrong?

Helen: I don't know. I feel like the Chinese internet is a hard thing to get a sense of from the outside, because it is so walled off. So,

this is definitely an area that's an example where I kind of had a broad idea going in and I now just have a much more sort of concrete day-to-day impression of what is it like to use WeChat all day, every day, for literally everything you do, which is something that almost all Chinese people do.

Or, which companies are people talking about, are they excited about, or not excited about? That's very much where I got this impression of Baidu, as opposed to there being sort of concrete, verifiable facts, if that makes sense.

Julia: Got it. Did you talk to your friends in the machine learning program -- or any other Chinese people you got to know -- about the social credit score phenomenon? That seems like another thing where there might be a lot of misconceptions, where it might be better or worse than we in the West think.

Helen: Yeah. That's actually a great example of an area where I feel like I can't say that that was something that I changed my mind about, because it just wasn't receiving much attention before I went to China, but it really, the story really blew up kind of during 2018. I do think it's one where the reporting in the West has been pretty sort of overblown and misleading.

It's actually, I find it interesting to compare two stories, if you like. Two of the biggest stories about China over the past year or so have been the social credit system story, and also the Uighur imprisonment and the oppression of the Uighur minority in Xinjiang, in China's far west. For me, that's an interesting contrast.

So in the social credit system side, it's tricky because it's not that the reporting has been drastically factually wrong. It's just been kind of misleading in a whole bunch of ways.

So, the story, the picture that most of my friends have who haven't spent any time in China, is that the Chinese government is rolling out on a massive scale this system that is going to look at every single aspect of your life -- who you're friends with on WeChat, who you message about what, what you're buying, how you're spending your time, how many kids you have -- and give

you a single unified numerical score. That score will determine all kinds of things about your life. For example, oh, my goodness, did you know some people are banned from taking high-speed trains or planes?

Is that kind of a maybe slightly dramatized version of the picture you have?

Julia: Not dramatized, no.

Helen: So, this is kind of pulling together and mashing together all of these different threads. Then, sort of hyperbolizing it because when you mash them together, it gets more scary and more Black Mirror-esque.

So, there's sort of two big things that are going on here. One is that China doesn't really have any kind of existing credit score system. Just straight up, how creditworthy is this person? Should we give them a loan?

So, there are now several commercial efforts to try and figure out ways of doing that, some of which do involve some pretty sketchy information. So, there's an app that Kai-Fu Lee, this Taiwanese-American venture capitalist, talks about, which you download onto your phone and you apply for a loan. It looks at all kinds of, basically all the data it can get from your phone including how much battery you have left, what model of phone you have and things like this, which seems like they shouldn't ... Maybe that's a little bit concerning as a way of determining if you should get a loan. But so, that's sort of the commercial side. So, that sometimes does involve having a numerical score, but that's not really that much different from the US credit score system, right?

Then, on the other side, you have this big government push. So social credit is definitely a big idea that the Chinese government is really interested in promoting and using. But this system of looking comprehensively at your whole life and giving you a single numerical score is really, as far as I can tell, not based in reality.

So, what's happening instead is that it's sort of -- and again, I should clarify. I don't want to sound like I think there's nothing concerning about this. I think there's plenty that's concerning about this. I just think that the way it's portrayed is so misleading that it's very frustrating to try and talk about it, because people immediately go to sort of the wrong place, if that makes sense.

Julia: Right. Yeah. I mean, after years of kind of following the media coverage of concerns about AI safety, just in the West, I'm already on a hair-trigger for misleading representations. So, you see enough images of the Terminator in response to much more nuanced positions, and you start to get suspicious.

Helen: Right. So, what the Chinese government is very interested in doing when it comes to social credit is thinking about how to apply existing laws in ways that are more suited to the digital age that we live in.

It is true that there have been a large number of Chinese citizens who have been barred from using high-speed trains or planes. But as far as I know, this is generally because it's much more of a thing where if you do something wrong, there is some clear sort of punishment or clear reaction to that specific wrongdoing that you did. That make sense?

So, in the high-speed train or plane situation, I think it's usually either because you've misbehaved on a high-speed train or because you have some unpaid debt to a court, for example. So you've been fined something, or you've gone to court and haven't paid your fees or something like this.

So, the idea then is, "Oh, well, if you are so poor that you can't pay your court fees, it seems like you surely can't afford these, like, expensive train and plane tickets," so you have to just buy the slower trains, which go to all the same places. It's just slower and cheaper.

And there are still systems that are doing things like using facial recognition to check if you're jaywalking, or automatically recognizing which cars are parked in the wrong places using

their license plates. But that last one is really not that different from stuff that happens in the West.

Again, I don't want to say that none of these are problematic. I think there's plenty about it, that's sort of concerning. But just the whole wrapping it up in this Black Mirror, like “the government gives you a number and that rules every aspect of your life” is just not at all an accurate picture.

Julia: Got it. Yeah.

Helen: Before I lose my train of thought, I'll just say it's been interesting because I started talking about social credit in comparison to the Xinjiang situation, with Uighurs.

Julia: Right. Yeah.

Helen: Something I found really valuable about my time in China... nine months is obviously way too little time to learn everything about a country, or gain a really deep understanding, but it did help me get really familiar with the community of Westerners. They sometimes call themselves “China Watchers,” who have made a whole career out of this, that spend a lot of time in China, speak much better Chinese than I ever will.

It was really interesting watching them repeatedly get irritated by how the social credit system was portrayed and how that was distorted and used for political purposes, in comparison to how they've reacted to the Muslim Uighur oppression situation, which is that that just seems to be reported completely accurately, as far as anyone can tell. It seems to be really a horrible situation that is, in fact, going on. So, that was sort of interesting as a sort of sociological exercise as well.

Julia: Exactly, yeah. Although I'm not glad to hear that their reporting on the internment camps is roughly accurate. That's not good news.

Another general misconception, or potential misconception, about China that I'm curious if you agree with, might just be the idea that we can know things about it with confidence. So, I've

been reading some warnings lately that our information about China is just far more unreliable than we realize, including everything. GDP figures, education, crime, population statistics. Is that your impression, too?

Helen: Yeah. I mean, I think my impression is that this is a bit of a debate that goes on within the China Watching community. I think there's rough consensus that any Chinese government data should be treated as at least a little bit suspicious. There's some people who know much more about it than I do who definitely think, for example, the GDP figures are really suspicious because they just show this continuing growth in a sort of way that becomes increasingly implausible the longer it continues.

Julia: I have really wondered about that, just because China is such a big part of the overall "decline in poverty" story that everyone I know lauds and shares on Twitter. I'm quite confident that there has been a large decline in poverty -- but the fact that China is such a big chunk of it, and the stats from China seem not totally trustworthy, just makes me worried.

Helen: I think that's almost a separate issue, actually. So, I think for me, the skepticism I've seen about Chinese GDP numbers sort of increases over the last 5 or 10 or maybe 15 years, of the sort of continuing growth. I think in terms of the reductions in poverty, there's both an economic story you can tell for that and there's also just the lived experience of people who spend time in China and who got to go to China in, say, the 70s or the early 80s, and then go back to China now and see these massive, massive differences in prosperity levels.

Julia: Yeah, but we don't know how big it is, right? Clearly there has been a large decline in poverty but we just don't ...

Helen: Sure.

Julia: Doesn't it seem like there could be a wide range of magnitudes...?

Helen: Yeah. Sure. I think in terms of "how could China be such a large proportion of the range," it really is notable how completely the Chinese Communist Party just totally ruined their own economy in the 50s and 60s.

So, there's a strong case to be made, or a strong sort of causal story to be told, for how there could be such giant growth just because they're starting from such a sort of self-inflicted point of great weakness, if that makes sense.

I do think that it does make sense to have general skepticism about any numbers coming out of the Chinese government or, to some extent also, other sources in China.

Julia: Got it. In your conversations in particular with the AI scientists who you got to meet in China, what did you notice? Did anything surprise you? Were their views different in any systematic way from the American AI scientists you'd talked to?

Helen: Yeah. So, I should definitely caveat that this was a small number of conversations. It was maybe sort of five conversations of any decent length.

Julia: Oh, you also went to at least one AI conference I know in China.

Helen: Yes, that's true. But that was much more difficult to have sort of substantive, in-depth conversations there.

I think a thing that I noticed in general among these conversations with more technical people... so in the West, in similar conversations that I've been a part of, there's often been a sort of part of the conversation that was dedicated to, "How do you think AI will affect society, what do you think are the important sort of potential risks or benefits," or whatever. And maybe I have my own views, and I sort of share those views. And usually the person doesn't 100% agree with me, and maybe they'll sort of provide a slightly different take or a totally different take, but they usually seem to have a reasonably well thought-through picture of "What does AI mean for society? What might be good or bad about it?"

The Chinese professors that I talked to -- and this could totally just be a matter of relationships, since they didn't feel comfortable with me, but -- they really didn't seem interested in engaging in that part of the conversation. They sort of seemed to want to say things like, "Oh, it's just going to be a really good tool, so we'll just do what humanity or just do what its users will want it to do, and that's sort of all." I would kind of ask about risks, and they would say, "Oh, it's not really something that I've thought about."

There's sort of an easy story you could tell there, which might be correct, which is basically: Chinese people are taught from a very young age that they should not have, or that it's dangerous to have, strong opinions about how the world should be, and how society should be, and that the important thing is just to fall in line and do your job.

So, that's one possibility for what's going on. Of course, I might have just had selection bias, or they might have thought that I was this strange foreigner asking them strange questions and they didn't want to say anything. Who knows?

Julia: Well, I mean, another possible story might just be that the sources of the discourse around AI risk in the West just haven't permeated China. Like there's this whole discourse that got signal boosted with Elon Musk and so on. So, there's been all these conversations in our part of the world that just maybe aren't happening there.

Helen: Sure, but I feel like plenty of the conversations I'm thinking of in the West happened before that was so widespread, and often the pushback would be something along the lines of, "Oh, no. Those kinds of worries are not reasonable. But I am really worried about employment and, like, here's how I think it's going to affect employment," or things along those lines. And that just didn't come up in any of these conversations, which I found a little bit surprising.

Julia: Sure. Okay, so you got back from China recently. You became the director of strategy for CSET, Center for Security and

Emerging Technology. Can you tell us a little bit about why CSET was founded and what you're currently working on?

Helen: Yeah. So, we were basically founded because -- so our name, Center for Security in Emerging Technology, gives us some ability to be broad in what kinds of emerging technologies we focus on. For the first at least two years, we're planning to focus on AI and advanced computing. That may well end up being more than two years, depending on how things play out.

The reason we were founded is essentially because of seeing this gap in the supply and demand in DC, or the appetite, for analysis and information on how the US government should be thinking about AI. In all kinds of different ways.

So, the one that we wanted to focus in on was the security or national security dimensions of that. Because we think that they're so important, and we think that missteps there could be really damaging. Yeah, so that's sort of the basic overview of CSET.

Julia: So, it sounds like the reason that you decided to focus on AI specifically out of all possible emerging technologies is just because the supply and demand gap was especially large there for information?

Helen: That's right. That's right.

So, what we work on in the future will similarly be determined by that. So, certainly on a scale of 10 or 20 years, I wouldn't want to be starting an organization that was definitely going to be working on AI for that length of time. So, depending on how things play out, we have room to move into different technologies where the government could use more in-depth analysis than it has time or resources to pursue.

Julia: Great. And when you talk about AI, are you more interested in specialized AI -- like the kind of things that are already in progress, like deep fakes or drones? Or are you more interested in the longer-term potential for general superintelligence?

Helen: Our big input into what we work on is what policymakers and other decision-makers in the government would find most useful. So, that kind of necessarily means that we focus a lot on technologies that are currently in play or might be in play in sort of the foreseeable future. More speculative technologies can certainly come into our work if we think that that's relevant or important, but it's not our bread and butter.

Julia: I saw that one of the main areas of interest at CSET is how AI interacts with other technologies. Can you give an example of that?

Helen: Yeah. I mean, there are several. So, a couple of obvious important ones would be how AI interacts with nuclear technology. So, this could have, again, several branches. So, it could be how does AI interact with the nuclear command and control line? So should we worried about cyber attacks on nuclear command and control, or does that not matter? Is it all air-gapped safely? I don't know the details of that.

Or you could be interested in AI's effects on nuclear deterrence, and is that going to change this extremely unusually stable balance we've had at the international level, for the past half-century or so.

Another area that where AI overlaps with existing technologies is in this cybersecurity area, or cyber operations. So this is how, for example, nation states, and not only nation states, basically hack each other. And to the extent that we think that AI or machine learning or reinforcement learning, or what have you, is going to make it possible to attack or defend computer systems in new ways... That could be relevant for that space, for example.

Julia: Great. In your interaction so far with American policymakers about AI, has anything surprised you about their views? Have there been any key disagreements that you find you have with the US policy community?

Helen: I mean, I think an interesting thing about being in DC is just that everyone here, or so many people here, especially people in

government, have so little time to think about so many issues. And there's so much going on and they have to try and keep their heads wrapped around it. This means that kind of inevitably, sort of simple versions of important ideas can be very powerful and can get sort of really stuck in people's minds.

I see a few of these that I kind of disagree with, and I kind of understand why they got an embedded -- but if I had my druthers, I would embed a slightly different idea.

So, an example of this would be, in terms of AI, the idea that data is this super, super important input to machine learning systems. It's sort of step one of the argument. Step two of the argument is: And China has a larger population and weaker privacy controls, so Chinese companies and Chinese government will have access to more data. That's step two. Therefore, conclusion: China has this intrinsic advantage in AI.

Julia: Right. Yeah. I've heard that framed in terms of the metaphor where data is like oil.

Helen: Right. Exactly.

Julia: So, China has this natural resource that will make it more powerful.

Helen: Exactly. And, again, this is -- each step of the argument is not completely false. So, certainly data is important for many types of machine learning systems, though not all. Certainly, China does have a larger population and it does seem to have weaker privacy controls in some ways, though not in others.

Actually, an interesting comparison between China and the US seems to be that US or American citizens are very concerned about their privacy from the government, but are much more willing for companies to have access to their data, if it means they can get better products or whatever. Whereas Chinese citizens are much more concerned about whether companies can access their data, and are much happier with the government sort of having access.

Julia: That’s so interesting, do you know why?

Helen: I don't really. You can tell a story about the US, in terms of its attitudes toward government here and so on. Similarly, you can go back to the same story about China, with regard to the government being sort of a big player in the massive boom in prosperity that they've had. But I can't get more detailed than that. I don't really have evidence for that.

Julia: Makes sense. Great. So, why is the story flawed?

Helen: Yeah, so I think firstly, I think the first argument about the importance of data for machine learning overstates how important data is, and certainly overstates how monolithic data is.

I guess that maybe gets to the second argument, which is about the key fact here is how many people you have in your country, and how easily you can access data about those people. Which, I mean, I think it's tricky. Because the argument is often made with the sort of hand wavy, scary face, at the end, of like, “...Bad things happen."

But if you're looking at what bad things might happen, it seems like plenty of other types of data where the US has a huge advantage:

Anything to do with military data, whether it be satellite imagery, or data from other sensors that the US government has, the US is just really going to have a big advantage.

The whole internet is in English.

From what I've read, self-driving car input data tends to be much stronger in the US than in China.

There's just many, many types of relevant data. And what's relevant for any given machine learning system will be different from any other, depending on application. So, just to go from consumer data, to all data, seems like it misses a lot. Aside from the whole question of how do the privacy controls actually work,

and how well can Chinese companies actually integrate data from different sources, and so on.

Julia: Right, right. No. That's a good point.

I read somewhere that in addition, it seems like data has steeply diminishing marginal returns. So, China might even just have much less of an advantage, even setting aside the factors you mentioned, than people think. Does that sound right to you?

Helen: Yeah. I'm not sure that I have a strong enough grasp of the machine learning state of the art understanding of this to confidently say. I wouldn't be surprised if it was the case that data has a diminishing marginal returns, if you have sort of a static amount of hardware that you're applying to training. But that, if you continue to get value from data, if you can add more hardware, that the story might be a little more complicated. But I'm not sure.

Julia: Got it. Yeah. So, another part of this "data as oil" metaphor is that insights and expertise from programmers is, I guess, less important. Is that another part of the story you think is flawed?

Helen: Yeah. I think that one is more something that we're just pretty uncertain about. That's sort of inherently making a forecast about: How interesting are advances in machine learning going to be, and where will they come from?

Someone who's very prominently associated with this and has made the argument in his book is the same Taiwanese-American venture capitalist I mentioned earlier, Kai-Fu Lee. He's also a former AI researcher, I should say, as well, so he knows his stuff. But he has made the claim that sort of "returns to ingenuity" are going down at this point. And that it's really just about sort of grunt work, and that China has a big advantage in putting in the grunt work. Is a highly-condensed version of his argument.

I just don't really see that. It seems to me like things like the recent StarCraft result from DeepMind, where they had their algorithm -- or had their system, rather -- beat top professional

players. It was a somewhat restricted version of the game, but it was not extremely restricted.

Or this release from OpenAI, which put out GPT-2, which was a text-generating system where you sort of give it a prompt. You give it, say, the first sentence or the first paragraph. It can then generate this long text, sort of however long you like, but it's really much more plausible and much more human-like than anything we'd seen before.

They seem like not trivial advances to me. And they both came out of two of the labs that are known for having the very best people in the world. So that says to me that the "returns to excellent people" seem to still be there, as far as I can tell. But predicting the future is sort of hard to say.

Julia: Going back for a moment to the US government and their thinking about AI: It has seemed to me that the US government has not been very agent-y when it comes to anticipating the importance of AI. And by agent-y, I mean like planning ahead, taking proactive steps in advance to pursue your goals. Is that your view as well?

Helen: I think, again, with having moved to DC and getting used to things here and so on, it seems like that's kind of true of the US government on basically all fronts. I'm not sure if you disagree. It's a gigantic bureaucracy that was designed 250 years ago. That's not completely true, but the blueprints were created 250 years ago. It's just enormous, and has a huge number of rules and regulations and different people and different agencies and other bodies with different incentives and different plans and different goals.

So, I think, to me, it's more like it's kind of a miracle when the US government can be agent-y about something. Then…I feel like it's kind of not fair to expect anything else of a structure that is the way that it is.

Julia: Have we been getting more agent-y in any significant ways?

Helen: Oh! That's a really interesting question. I would have to ponder that and get back to you.

Julia: Well, I mean, one thing I'm interested in is I had heard, I don't know, at least a couple years ago now that the US government doesn't have very many people who have both technical AI knowledge and also who speak Chinese. Which seems like a big gap. Do you know if there are any attempts to fill that gap?

Helen: Yeah. So, it's a tough thing to comment on because I think the people who know this for sure usually know it by having a security clearance and having access to classified information. It certainly does seem like the process that we have in place for how clearances are done makes it very difficult to have spent a long time in China, certainly extremely difficult to go back to China while you hold your clearance. Which seems like it would push in this direction but I don't feel comfortable making any blanket statements, just because it's hard to say who is and isn't employed.

Julia: Sure. That's fair.

Helen: Maybe one example of the US government trying to become a little bit more agent-y actually is an interesting case. You probably heard of the Project Maven situation.

Julia: Yeah.

Helen: So, that was actually ...

Julia: You want to just explain briefly what that is?

Helen: Sure, sure. So, basically Project Maven was a project set up within the Department of Defense where the whole motivation was sort of: AI is going to be really important. We need to start using it. Our existing procurement processes are not at all designed for anything remotely software-related. They take years and years and involve this contractor bidding process where even becoming a contractor involves a huge amount of annoying and slow paperwork. So, why don't we try a totally

different setup that is much more, is supposed to be inspired by the sort of Silicon Valley agile-type setup.

So, Project Maven was very deliberately designed to be, "Let's set up a project within the Department of Defense that uses machine learning, that uses a kind of commercially-mature application of machine learning."

They chose Computer Vision. I think as far as I know, they deliberately chose an application that didn't involve killing people. It was analyzing imagery, basically. Because the Department of Defense has just this huge amount of imagery that it collects from satellites and drones and other sensors. I think in this case, it was drone imagery that was being analyzed.

Employing people to sort of scan through that imagery and look for important things, interesting buildings or interesting whatever, military bases or other things that one might want to see in those images -- it takes a lot of manpower and a lot of time and a lot of money, and they still can't get through all of it.

So, they thought, "Great. This is an application of machine learning. The technology exists. It's not speculative. It doesn't involve killing people. Perfect trial case."

I was giving that as an example of a relatively agent-y thing that I think the Pentagon put together and got going on. Of course, the reason that most people are familiar with it is that Google, after being somewhat involved for a while, faced a lot of employee pressure not to work with the Department of Defense on it. I think largely because it involved drones, and the drone program is obviously pretty unpopular in many circles that overlap heavily with sort of Google employee circles. So, Google ended up pulling out of that project.

So I don't know. Maybe it wasn't so agent-y after all, but I thought it was-

Julia: Well, I think we can still give them credit. Yeah, give them a couple agent points.

I saw that another one of CSET's planned areas of focus is researching “which measures provide the clearest view of AI capabilities in different countries.” I'm quoting the founding director there, Jason Matheny.

So, I'm curious -- one thing that people often cite is that China publishes more papers on deep learning than the US does. Deep learning, maybe we explained that already, it's the dominant paradigm in AI that's generating a lot of powerful results.

Helen: Mm-hmm.

Julia: So, would you consider that, “number of papers published on deep learning,” would you consider that a meaningful metric?

Helen: I mean, I think it's meaningful. I don't think it is the be-all and end-all metric. I think it contains some information. I think the thing I find frustrating about how central that metric has been is that usually it's mentioned with no sort of accompanying ... I don't know. This is a very Rationally Speaking thing to say, so I'm glad I'm on this podcast and not another one...

But it's always mentioned without sort of any kind of caveats or any kind of context. For example, how are we counting Chinese versus non-Chinese papers? Because often, it seems to be just doing it via, "Is their last name Chinese," which seems like it really is going to miscount.

Julia: Oh, wow! There are a bunch of people with Chinese last names working at American AI companies.

Helen: Correct, many of whom are American citizens. So, I think I've definitely seen at least some measures that do that wrong, which seems just completely absurd. But then there's also, if you have a Chinese citizen working in an American university, how should that be counted? Is that a win for the university or is it win for China? It's very unclear.

And they also, these counts of papers have a hard time sort of saying anything about the quality of the papers involved. You

can look at citations, but that's not a perfect metric. But it's better, for sure.

And then, lastly, they rarely say anything about the different incentives that Chinese and non-Chinese academics face in publishing. So, actually my partner is, he's a chemistry PhD student, and he's currently spending a month in Shanghai. He mentioned to me spontaneously that it's clear to him or maybe it got mentioned explicitly that his professor's salary is dependent on how many papers come out of his lab. So that's just a super different setup. Obviously, in the US, we have plenty of maybe exaggerated incentives for academics to publish papers, but I feel like that's another level.

Julia: It is. Yeah. I don't know if you know this, but is the salary just a function of the number of publications? Or is it ... In the tenure system, at least in theory, they care about the quality of the journal they're publishing.

Helen: I don't know. I assume it's some kind of bonus and I have no idea how they account for quality.

Julia: Huh. I heard about a café in, I think it was Beijing, that gives you free meals for every paper that you get -- oh, I remember what it was. They take a discount off of your meal proportional to the impact factor of your last paper that was published.

Helen: That's amazing.

Julia: Many interesting incentive systems.

Helen: Yeah. Talk about distorting incentives.

Julia: So, a few minutes ago, we were talking about some of the flaws in the metaphor of “data as oil.” There's an even bigger metaphor, or framing device, that you hear in the discussion of AI in China and that is an “arms race” in AI. You've talked a little bit about and written about why thinking of technological development in AI as an arms race is kind of flawed. Can you say more about that?

Helen: Yeah. I mean, I think there's a few ... I have a few different issues with this framing so I can try attacking it from a few different directions.

Julia: Great.

Helen: I think one important difference here is just that AI is not a single technology. It's a sort of underlying type of algorithm, or something like this, that can power many, many, many different types of applications in many, many, many different ways.

I think a lot of American thinking on defense strategy and the geopolitical order and so on, is kind of naturally inspired by the Cold War, and by the models that people are used to thinking in, based on our recent history. Nuclear weapons were a really massive feature of the Cold War. There you could very much ask the questions of: Does this country have nuclear weapons? How many do they have? What kinds? You can add them up.

That just really doesn't work with AI because it's so ... Firstly, because even in the military domain, I think it's much more like electricity, in that it will sort of eventually power and sort of seep through all possible domains, not just powering actual weapons, but also in logistics and planning and transportation and all of these other domains. So, it's not really something ... It's going to be much more difficult to sort of easily compare capabilities, even if you purely restrict it to the military domain.

But then, of course, it also is not really reasonable to think about it as an arms race because AI is so, as I think Jack Clark, the policy director of OpenAI, has termed, or at least popularized the phrase “omni-use.” So, it can be used in just every possible domain. Again, here I think the electricity analogy while, of course, not perfect, I think is pretty good here as sort of giving a sense of how broadly these technologies could spread.

So, any technology that can be used so broadly is surely going to have some applications that have this sort of relative dimension, of who is better, or who has more. For example, in number of nukes.

But it's also going to have these massive non-zero-sum components. For example, if Google builds some new, I don't know, the Google assistant or whatever, Apple's Siri, and uses AI to make that better, that's just sort of going to be a boon to consumers around the world. In a very absolute sense, and not at all relative. So, that's certainly one big way in which I find the arms race framing misleading.

Julia: Something I'm confused about with the arms race framing, especially when it's US versus China, which is the usual context, is that most of the major AI development is happening in private companies, right? Not government.

Helen: Mm-hmm.

Julia: Which is, again, a disanalogy with the Cold War. So, how does it even help the US geopolitically if an American company is developing powerful AI?

Helen: Yeah. So I think this is an area that we would really like to dig into at CSET, actually, because I think it's really interesting. I think it is relevant how domestic industry is doing. I think that is ... Sorry. To finish the thought, I think that is relevant to kind of military strength or hard power, as it might get called.

Sometimes, it gets contrasted between hard power and soft power, where soft power is more the cultural and fuzzier side of things, influence-based side of things. So, as well as talking about arms races, a really common framing that I hear as well is talking about competitiveness or competition.

I think another thing that's sort of going on there is that there's this underlying concern about the rise of China for reasons that have very little to do with AI really, just the overall macroeconomic shifts that we're seeing, that China is becoming larger and more powerful. There's a lot of anxiety about that, and about that displacing the United States's kind of unusual position it's held for the last 30 or so years as basically the only world power.

So, I think AI has sort of stepped in, in this moment, and that has made it possible for many of those anxieties to be mapped onto AI as a technology. And for AI to kind of bear the mantle of... As Putin famously said -- though he actually said this to school children in a science fair, trying to encourage them -- "You know, whoever rules in AI will rule the world," or whatever.

Julia: It's so funny how that quote in a very innocuous situation got turned into a big thing. I mean, he may well have meant it but ...

Helen: Right, but I think it really does express perfectly this sort of mapping of overall concerns about geopolitical balance onto AI as this single technology that is the be-all and end-all.

Julia: Got it, yeah. So, it's just a much broader and fuzzier notion of "arms race" that includes economic strength and other things, too.

Helen: Right. And again, I don't want to make it sound like I think there's nothing to be concerned about here. I think there's plenty. For example, I think the Chinese government is extremely authoritarian and is going to use these technologies to cement that and will be perfectly happy to sell those technologies, then, to other countries if they'll pay them money to also use them on their own populations. I think that's extremely concerning. I just think it's different from sort of "AI's this one technology, and you just have to be best at AI, and then you'll win, in some sort of extremely undefined sense of winning."

Julia: Are there any historical situations that strike you as being more usefully analogous to the development of AI than the nuclear arms race was?

Helen: Yeah. I mean, I'm interested to dig more into this electricity analogy because I think it's a pretty good one. I think it also, an implication I've been thinking about a little bit, especially as I've been sort of moving in in defense-related circles is to think in terms of if you're interested in how electricity can affect your military or improve your military, you're going to have to do

this very wide ranging sort of completely rebuilding your infrastructure to make it compatible with how electricity works, right? I think there's sort of a similar thing that could be said about AI. I think a big thing that, if the Department of Defense, for example, is really serious about implementing AI, the first thing it's going to need to do is just improve all of its digital systems, which are extremely outdated and haven't been invested in.

So, I think the electricity example is a pretty good one or I think you can get some juice out of it. I don't know, the other one that sort of is often tossed around that I think is not terrible is just the industrial revolution as a whole, though it's a little bit less clear there what AI is. Like, is AI the same as the steam engine?

Julia: Yeah.

Helen: Something along those lines.

Julia: It's tough. It's tough finding good mappings and analogies that don't fall apart. I mean, I guess I was thinking more specifically about the geopolitical ramifications of AI development. And the game theory in the Cold War was so simple. A simple game theory never maps super well onto the real world but, still, it was kind of a clear framework to use to think about strategic considerations.

I'm just wondering if we have anything like that for our current very messy situation now, which, as you've described, has many different kinds of AI for different situations and it has companies and governments and so on and so forth.

Helen: Yeah. I know. Again, it's the best thing I can come up with in the moment is the electricity analogy, or the changes in technology that were happening at the start of the 20th century, roughly. I think an interesting ... Something that I know some people are concerned about with AI is the risk of unintended escalation, right?

Julia: Mm-hmm.

Helen: So, perhaps you have some kind of automated systems in some kind of battlefield context and they interact with each other in an unexpected way, and escalate a situation in a way that is not what the humans involved intended. I think that's interestingly analogous to at least one story that I've heard about how the First World War got started. Are you familiar with this one?

Julia: Oh, why don't you tell it?

Helen: It's basically around these various European nations having these modernized or somewhat modernized militaries that they haven't really fought in wars with before, and there being all these new considerations. The railroad I think was also quite new.

So, there are all these new considerations about, if you start mobilizing your troops on the train at this time, what does that imply for when the other country needs to have already mobilized its troops before, in order to be ready? This sort of ends up with this strange dilemma, where everyone kind of needed to start preparing before anyone actually really wanted to go to war -- ending up, again, with this sort of similar unintended escalation dynamic.

I'm not a historian so I don't want to stand 100% behind that causal explanation, but it's kind of interesting that it has that neat analogy, again.

Julia: Yeah. It really does. I have to admit that, personally, I've been feeling kind of pessimistic about the potential for cooperation around AI development, especially between countries but also between companies. So, I'm hoping you can help me. I'll give you a couple reasons for pessimism and then you can share your thoughts and hopefully, counter some of my pessimism.

The first reason is just that... you've probably heard this statement bandied about, which is that “AI is software and it's impossible to regulate or control software.” So that makes any kind of treaty that relies on observation, like the, was it the Montreal, or Kyoto Protocol about climate change? That, you can kind of observe whether other people are adhering to the

treaties. It's just much, much harder when the technology is so much more invisible, as AI.

Helen: Sure. Yeah. I mean, I think that is right about software. I think actually it does look like AI will be very difficult to monitor in that way. I know that Miles Brundage, a friend and colleague of mine who works at OpenAI, which is a San Francisco-based AI research organization that I've mentioned a couple of times without introducing it.

So, Miles Brundage is on their policy team. He is extremely interested in ways in which, for example, AI hardware – so, the chips involved -- could be used as some kind of input that could be monitored, in the same way that, for example, uranium is closely monitored in the nuclear case.

I think it's a difficult question. It's not obvious how you could do this, but that is a much more sort of concrete, trackable thing that might be possible to build treaties around, or something like this, if at some point in the future, that seemed like something we wanted to do.

Julia: All right. Reason for pessimism number two is that cooperation, especially in this case, I think, it just isn't very robust. Even if 9 out of 10 major players in AI want to cooperate, it kind of still doesn't work if you don't have that 10 out of 10. “Doesn’t work” in the sense of the 9 out of 10 may not be willing to cooperate unless you can get the 10th, and also “doesn't work” in the sense that if the 10th player goes on and develops some powerful and unsafe system, then we all suffer, presumably.

Helen: Yeah. I don't know. I guess I kind of want to ask at this point if there's a more concrete version of cooperation that you're thinking about because I feel like my response to that would depend on the specific way in which different actors are trying to cooperate.

Julia: Okay. Yeah. I was lumping together a lot of different things in there. So, I was thinking about if there was a set of safety standards that everyone was going to agree to adhere to, like

intermittent testing of their AI systems in kind of constrained environments...

Helen: Yeah. I think you're probably right that ... I think in general, if you have most situations where you need kind of 10 different players to agree to something, to cooperate in a prisoner's dilemma, for example. That's just going to be a heavy lift.

Julia: Yeah. That's kind of what I'm picturing.

Helen: Sure.

Julia: I mean, it's not exactly a prisoner's dilemma but ...

Helen: Right, and that's why I pushed back a little bit because I think it does depend on what exactly is the situation, and what exactly are you asking the actors to do. I mean, maybe this is just something that I've kind of ... A reason that I've kind of turned away from talking about "cooperation" in this broad sense. Because it's just not clear to me that it's that helpful as a sort of overall category of actions that people could take.

It seems to me like, for example, investing in safety is a still fairly abstract but pretty concrete action that companies or universities or whatever could take. That seems like something where there is some amount of coordination needed. If you really felt like you were going to be missing out by investing in safety rather than just pressing full speed ahead, that would be a tough situation.

But I don't know. In real life, it's a bit more complicated than that. [There are] several different subfields that you could lump together under "safety," so things like interpretability, how easy is it to understand what a machine learning system is doing and why? Or robustness and security, how easily can you trick this system, how well would it work in settings where it wasn't designed for? Or things like value learning, how do you ask a machine learning system to optimize for something that is as complex and nuanced as human values?

Those subfields have become reasonably well-established as respected, normal machine learning research to do. So now it's not necessarily such a cost if a lab has a wing of people thinking about those problems, and then publishing their results for anyone else to use. That doesn't have very much of a prisoner's dilemma dynamic to it. So, we could totally end up in a more prisoner's dilemma-like situation in the future, but I don't think that's obviously that we will.

Julia: Okay. Interesting. Yeah, so I'm now feeling a little bit more optimistic that there might be avenues or solutions that weren't salient to me before, that involve reducing the costs of cooperation and requiring less, kind of, trust.

Helen: Glad I could be cheery.

Julia: Yeah. Yeah. Do you want to add any additional reasons for optimism?

Helen: I mean, I think at a very high level, I think it's kind of nice the machine learning community is so international and open and committed to trying to do good things. I think there could be more where that came from, so I think there could be more connection between Chinese and American researchers, or Chinese and Western researchers in general. I think there could be more thought put into what exactly the good things that machine learning could do for the world are, and how researchers can promote those

But it seems like we're starting from a pretty good baseline here. Seems like there are plenty of fields that would be much worse on those dimensions. I think that's the main additional reason that comes to mind. I'm not sure, to be clear, I'm not sure, in balance, if I feel optimistic or pessimistic about cooperation. I feel like I haven't-

Julia: I did kind of slot you into the optimistic spot in this conversation, so that's not your fault, but yeah. I mean, if it were super easy, we wouldn't really need a CSET so that makes sense.

Helen: Right.

Julia: Before we move onto the Rationally Speaking picks, I wanted to ask: Is CSET still hiring? Because, if so, we should put in a plug for that.

Helen: Oh, yeah. We are hiring. We're starting to slow down our hiring, so by the time this airs, we may just have a thing on our website saying that you can send us your resume, but please do. We may also still be hiring in full tilt.

So we're hiring for research fellows to lead our research projects, hiring for research analysts, which is a slightly more junior role, but also looking for ... We'd love to have at least one person on staff who really knows AI and machine learning really, really well. So, we have a post for an AI and machine learning fellow on our website. We're also hiring for a data scientist, a senior software engineer. I think that's basically all the roles at this point.

Julia: Okay. Great. The website address is?

Helen: The website is CSET, so C-S-E-T @georgetown.edu.

Julia: Great. So, we'll add that link on the podcast website as well.

Now, Helen, before I let you go, do you have any recommendations for our listeners for ways to keep abreast of developments in China, or ways to get background on China, resources that you think are particularly trustworthy or interesting?

Helen: Yeah. I have a couple of books that I'd love to recommend, which are part of the sort of series of books I read and other resources I looked into as I was moving to China. So the two that I enjoyed the most, one is called “The Beautiful Country and the Middle Kingdom,” by John Pomfret. That's the translations of America and China in Chinese. That's just a broad history of the US-China relationship, starting in the 18th century and up to today, and is very comprehensive and interesting.

Then, the other is called “Age of Ambition,” by Evan Osnos. It's also non-fiction, but it's just a sort of a potpourri of stories of modern Chinese people and what their lives are like, and how they think about, what they want out of life and so on. So, I found those two provided a really nice, sort of... one, this broad background overview, and another that's sort of a colorful set of pictures of what life is like for different kinds of people.

Julia: Excellent. That's a really good pair. Well, Helen, thank you so much for being on Rationally Speaking. This was very enlightening.

Helen: Thanks. I had a great time.

Julia: This concludes another episode of Rationally Speaking. Join us next time for more explorations on the borderlands between reason and nonsense.