Accelerating AI Ethics

AI and Democracy: Ambassador Audrey Tang on Plurality in Practice, Transparency and Collective Intelligence

Transcript

July 2025

www. afp.oxford-aiethics.ox.ac.uk

00:00:00 Green

Hello and welcome to the accelerating AI ethics podcast of the University of Oxford. I'm doctor Caroline Green, and in each episode, we explore bold ideas, innovative thinking, and creative responses to the ethical challenges posed by artificial intelligence.

Today's guest is someone who truly redefines what's possible when technology meets democracy.

Ambassador Audrey Tang is Taiwan cyber ambassador at large, and formerly its first digital minister. A self-described civic hacker Audrey helped transform Taiwan's government into a global leader in digital democracy, using open data, participatory platforms and radical transparency to foster public trust.

Now, as a fellow of the Accelerator Fellowship programme here at Oxford, she is developing a project called '*Plurality Advancing Ethical AI through Collective Intelligence*', a bold vision, and one that offers new ways of thinking about how societies can shape AI, not the other way.

Audrey, welcome. It's a real pleasure to have you with us today.

00:01:07 Tang

Very happy to be here. My first day as an accelerator fellow, really good to be in this new space with this podcast and sharing with you all, on how to align AI to society, and all the other way around.

00:01:18 Green

To get started, tell me about Audrey Tang. Audrey, where are you from? What's your story?

00:01:26 Tang

So I was born in Taipei, Taiwan. When I was four the doctors told me and my family that this child has a heart defect and has only a 50% chance to live until surgery. I eventually got surgery when I was 12, but for the first 12 years of my life I go into sleep not knowing whether I'll wake up. Feels like flipping a coin. And so this gives me two kinds of superpowers. One is that I learned Daoist meditation, very early on, so that whenever my heart beat above a certain beats per minute, I start deep breathing and so on. Instinctively, because if I don't, I just faint. And the other thing is that I publish before I perish. So, every day I record what I learned that day, first in tape recorders. You know those plastic things? And eventually, of course, to the Internet. So, I got into the habit of publishing everything into the public domain, into Creative Commons, because if I don't wake up the next day, well, people are going to be able to use it and remix.

00:02:27 Green

Wow, that's that's quite a story. So as a child, when you were going through that, you actually took it as a driver to, you know, share with the world, what you're learning every single day.

00:02:40 Tang

Right. It's an existential opportunity.

00:02:43 Green

And how has that shaped you into adulthood?

00:02:46 Tang

So, when I was 14, I already went through three kindergartens, six primary schools, and one year of middle school. And I discovered at a time this thing called the Internet web browsers, which was just coming up in 1995 and also this preprint server called arXiv, from Cornell, it is still around, and is where most people publish their AI findings and so on. And so, I started writing professors and they didn't know I was just 14. I actually spoke terrible English back then, I had to look up dictionaries, but I was very interested in why people come to trust each other so readily online; in person it takes hours, days for people to get acquainted to one another. But online with just the right meme, the right hashtag, the right common ideas, the right domain name, people tell each other very intimate things very quickly. And of course, they also break up very quickly, but that's called Swift Trust.

And I made it one of my research topics. I did the science fair topic in the nation, and I started learning about this research network. People are looking into this kind of new emerging social phenomenon on the Internet. And I told the head of my school, Principal Duhai Ping, I said look, I can do 16 hours a day doing research on this new thing called the Internet or I can go to your school, pretend to study for 8 hours, and then only do 8 hours research. I would like your help and she read my e-mail printouts and say OK, from tomorrow you don't have to go to my school anymore.

00:04:26 Tang

And I was like, OK, but it's compulsory education. My family will get fined and she's like, it's OK. I'll just fake the records for you (and I tell this story because it's more than 20 years, the prosecutor period is over). But the point is that I learned that bureaucracy is actually innovative. People are willing to actually bend the rules. If you say that actually you're of the same value, but we're just taking a shortcut, a better way, more effective way to accelerate getting to that common purpose together.

00:04:56 Green

So, Audrey, listening to your amazing story, it strikes me that you are someone who is really fascinated by people, by humans, how their minds work. Can you tell me a little bit more what drives you working with humans?

00:05:16 Tang

That's a great question.

Indeed, when I quit school, I almost immediately put my learning into use by co-founding, one of the fastest growing taiwanese.com enterprises, a startup called [Emporium?], and we eventually got investment from Intel, started CoolBid which is the equivalent of eBay, like C2C auction and search engines. And what I've witnessed is that people are much nicer when they're around other people. If you pole them individually, if we put them into a place where they're isolated and just look at snippets of social media posts, enragement driven engagement and so on, then actually people become very social, on the other hand, if we put people in groups like a group of ten, and they understand what each other are saying, each of their typing is, within the shared context, then people start moving from 'IMBY or 'NIMBY, like very selfish positions into 'MIMBY', like 'maybe in my backyard', but only if you also commit something in your backyard. And so, I think the social nature of the Internet was always what fascinates me and how to design such spaces, such as in C2C auctions, that elicit the best from the people, the most reputable from people, the most ethical from people. That has been kind of my main thing as an entrepreneur.

00:06:45 Green

So, do you think that humans are better when they are operating collectively?

00:06:52 Tang

Yes, and this is what we call pro-social media in Taiwan. So, a very quick story. In 2014, we peacefully occupied our Parliament for three weeks in Taiwan, because the president at the time was enjoying only 9% approval rate. So, in the country of 24 million people, anything President Ma says 20 million people were automatically skeptical. Which was the same situation in many of the Occupy Arab Spring movements. But we took a critical different approach. Instead of calling ourselves protesters, which are against something, we call ourselves demonstrators, which is showing something new, and that's something new. Is anybody who are worried about the trade agreement with Beijing at the time that would have, you know, invited Huawei, ZTE into our telecommunication our 4G infrastructure our publishing industry, media and so on. If you're worried instead of protesting, saying we shouldn't do this, you can go to one of those corners in the occupied Parliament on the facilitated conversation where people ask each other how do I truly feel about this? And starting from the position of feelings, peoples cohere, now set of very coherent ideas. By the end of the three weeks and the speaker of the Parliament Jinping simply said, ‘OK, the people's ideas are better than our ideas. So you win, we will ratify this, go home’, and so, we became the very rare occupy because of this small scale facilitated, group based, conversations actually converged instead of diverged.

00:08:32 Green

Wow, I love that. So here you're also bringing in a whole new way of looking at democracy and how technology can help us today to forge new paths in global democracy, right. And bringing people together. What I also really enjoy about what you're saying is you believe in the good in people, right? That's the whole way you're coming to your work, is that right?

00:09:01 Tang

Yeah, I believe people are good when they're around other people and so AGI to me is Augmented Group Intelligence. We should develop AI systems that augment our civic muscles instead of just – I talk to my chat bot, you talk to your chat bot, we send chat bots to deliberate, make decisions, persuade us. That's that's very impressive, but it's as impressive as me sending my robot to the gym to lift the weights, and you send your robot to run the treadmill. Very, very impressive. But our muscles don't grow this way. So, to me, this civic muscle, this relationship building between people who have different ideas, different ideologies, even, but managed to find a common ground –

I think that is core to democracy.

00:09:45 Green

So that's really interesting. I'd love to hear more about that, because quite often when I think of social media, when I think of artificial intelligence, it's more anti-human really, it's really anti positive relationships. You know we see a lot of bad behaviour on the Internet. We see echo chambers. So, your idea of actually bringing technology to the world to improve relationships between people to find innovative solutions. That's something really exciting. Can you tell me a little bit more about that. Are there specific projects you've been working on where you've seen the power of you know that?

00:10:28 Tang

Collaboration? Certainly so 10 years ago in 2015, many of the anti-social corner of social media become much worse than before, because they adopted this ‘for you’ algorithm. Prior to that people just followed each other and you got a chronological feed of what other people posted, that are your followers’ network, so people do have common experiences, common knowledge. But after the pivot to the ‘for you’ feed, they just figure out with the parasitic AI what keeps people addicted to the touch screens- and it turns out engagement is easier through enrichment. It turns out that people, when they're isolated in these small screens, prefer to see things that are much more sensational and polarising than what is healing or bridging. And so, it's like rewarding your children every time they are mean to somebody else. So, in this sense, this parasitic AI was the first misaligned AI system that turned a neutral platform into a non-ethical platform.

And in Taiwan, at the same time, we were experimenting with the other direction. We worked with the open-source system called Polis, but in 2015, we were trying the other direction, the pro-social media. We worked with Polis, an open-source system. When Uber came to Taiwan that year, many people are very afraid they would take the job of taxis, that this algorithmic dispatch is going to ignore the professional driver license system, and people would get maybe higher quality service, maybe lower quality service, really nobody knows. And we use the Polis system to basically ask the entire society ‘what do you feel about this situation?’ Because people are experts in their feelings. If we ask, ‘what do you think about sharing economy policy?’, very few people would be able to chime in, but because we just asked, ‘how do you feel?’ And we show people, once they share how they feel they can upvote, they can downvote on each other's feelings, but there's no reply button, so there's no room for trolls to grow, and we show people in their avatar which cluster they belong to – where the people are that share similar feelings. And that's good for two reasons. First, people understand that actually we do manage to agree with each other on some of those common feelings. For example, everybody felt that insurance is important. Everybody felt that while surge pricing is fine, but undercutting existing meters is not. And so instead of the polarising debate sharing economy, gig economy or whatever, it shows the connective tissue of this group. And the second is that there's a scoreboard; the longer distance your ideas resonate across groups, the more bonus you're given. The most viral ideas are the ones that are the most eclectic, that bring the uncommon ground, the rarely discovered common ground among people who initially diverged. After three weeks again we converged online so we didn't have to occupy any government buildings, but we replicated the same process we did in the Sunflower Movement and the top nine ideas that cross the threshold of 85% agreement in all the different groups, regardless of whether they are majority or minority. It unified the society together and we make the law about Uber using this rough consensus from the people. So over the next six years we would hold more than 100 of those collaborative meetings and the approval rate went from 9% to more than 70% by 2020.

00:14:14 Green

Wow, that's amazing. I'm wondering about when you tell these stories, how you bring people together, you know, collectively, on the Internet, it's very powerful in, in terms of beating social isolation, loneliness. But I'm also wondering about groups of people here who are digitally, not as connected. I work a lot with, you know, communities, individuals who don't use the Internet. They just don't have the connectivity or so on. What can we do to also reach these people?

00:14:51 Tang

Yeah. In Taiwan, broadband is a human right. We have the Universal Service Fund, so that the telecoms, they can go to even the top of Yushan almost 4000 meters high and set up broadband connections. If they cannot recover from subscription fees, because there's just fewer out there, the other telecoms who didn't make such investments must reimburse them on the cost lost. And so, the Universal Service Fund ensures that anywhere in Taiwan, no matter how remote the island, how high the mountain you have connectivity in the form of I think just £15 per month you get unlimited data connection and so we have been doing that for almost a decade now. And so because of that, no one is left behind. And for people who don't prefer to use the Internet, we use the strategy of 'Helping the Helpers. So, the young people in Taiwan, instead of just teaching them digital literacy, which would be about receiving information, we teach them digital competency. Competency is about finding these ideas together and, for example, when people have disagreement about air pollutions the young people set up air measurement stations. And then getting their parents grandparents to look at the numbers shared together, or they fact checked the three presidential candidates as they were having a debate and that they found some flaw in it, [then] maybe their name appears on national television and the younger people, younger than 18 were actually the most active on our national participation platform, so they would start a petition, for example, ‘saying let's go to the school one hour later because studies show one more hour of sleep is better than one more hour of study when it comes to grades’. And then they convinced the elderly citizens to help on their cause to help them to reach 5000 signatures and eventually got the time tempo changed. Of course, they don't just do things concerning themselves, and they also, for example petition for banning of plastic straw with bubble tea takeouts, and those petitioners become cabinet level advisors. Reverse mentors to minister become very famous. And then they inspire other young people to start even more ambitious things like starting a menstruation museum in Taipei and in just two years removed the taboo about this all together in all municipalities and so on. So with the young people as reverse mentors to senior people we ensure that even if you are in a rural place and you don't want to talk across the Internet with other people, you can talk to your young people who are like ambassadors to the digital world.

00:17:39 Green

Well, that's great. So, the idea of a human right to the broadband. And then that beautiful example of how technology can span the generations for intergenerational activity and bringing people together of all ages. That's very powerful thinking about, you know, our changing demographics. And so yeah, thank you for sharing that story.

00:18:08 Green

I'm going to pivot slightly to different questions. So, Audrey, what is AI ethics to you?

00:18:18 Tang

To me, AI ethics represents the technologies and methodologies to imbue the human society's preferences into the whole cycle of AI development so that we ensure that the AI systems that we make conform to the societal expectations, understand the social context and norms, and is steerable by the unity. Without AI ethics, it's like a car that is has a very fixed place to go, and the only thing you can do is to hit the brakes or hit the gas pedal. That is to say, to decelerate or to accelerate. On the other hand, most of the societies do not want a few people in Silicon Valley, or in some other big tech to dictate how we should relate to one another. So just like social media, many people would prefer the social media instead of selling our attention to the highest bidder, they would prefer that social media have more bridging content, more shared experience and so on. However, without ethics input, the social media network systems are designed in such a way that it pulls people, streamlines the social fabric, and then sell our attention to the highest bidder, and so had it been designed with ethics in mind they will come up with very different business models. Maybe instead of selling individualised advertisement, they will sell common experiences, curated experience or subscription-based business models and so on. To me, it's not just technical, but also about how the private sector incentive works.

00:20:05 Green

From Civic Hacker to Digital Minister to global advocate for Democratic technology, it's really extraordinary.

What drives you to build systems of radical transparency and public participation? We've already spoken about that now a bit, but just to put that question to you again, what's your main driver right now?

00:20:27 Tang

Sure, my main driver was the same as when I got into the cabinet position in 2016.

As I mentioned, the main problem we try to address is the deficit of trust. People were losing trust in all the vertical institutions, whether it's political parties or ministers or journalists or academic experts and so on. People much rather would like to trust at a time, people who sound like them, who look like them, who, you know, gets more comments. And so on. On the other hand, the social media, as we mentioned, was pulling those influences into extreme positions. So, we have two problems. One is that the government, the elites, just don't trust the people enough in order to win back trust. But to give no trust is to get no trust. So, the first thing we addressed with radical transparency is to radically trust the people; if the people see a public service that is not designed well, instead of protesting on the demand side, they can switch to the supply side by demonstrating how to do it better. If they feel that the contact tracing is not respecting privacy because of radical transparency, they can design better contact tracing system that preserved the privacy and indeed helped Taiwan to last until Omicron and we never locked down any cities and we reported one of the best economic growths during the three years. That is not because the elites have good ideas. That's because the people can co create together. And the other one is to depolarise, to bring people back from the extremes where people hate each other, especially intergenerationally. That was a big issue in Taiwan, so the young people, instead of just lamenting that they cannot outvote senior citizens, because of declining population, invite them on the table and then build intergenerational links. And so today, Taiwan is doing very well according to BTI, more than

90% of Taiwanese said that democracy is at least fairly good. On the other hand, no country is an island, not even Taiwan. So even if Taiwan depolarised the society, even if Taiwan rebuilt the democratic resilience if all our allies succumb to the polarisation, to the hate to the intergenerational distrust, then this authoritarian notion that democracy only leads to chaos, democracy never deliver is still kind of like self-fulfilling prophecy. So my main job now is to show that actually democracy can deliver and just with some tweaks in the social media regulations in the ethics, when we're designing AI, we can actually steer those technologies toward a prosocial direction, instead of this singularity vision where it just keeps getting better and better, ultimate its next generation without the need of humans in training, new AI models – superintelligence – take off – leaving everyone behind.

I think the world needs a better vision than this superintelligent singularity take off, and which is what I call Plurality.

00:23:39 Green

Yeah. So, what strikes me about that is that the Taiwanese government also has the readiness and the openness to listen to people, to embrace that modern age. Democracy. Is this something you feel we see or you can see also in other governments around the world?

00:24:01 Tang

Definitely so, I would say first that we learned this technique from many smaller polities. Our e-petition system was from Iceland *Better Reykjavik*, our participatory budget system are from Porto Alegre in Brazil as well as from Madrid and Barcelona (Consul and Decidim), respectively, the Polis system, was from Seattle, the [Lumio?] system, which we used during the Occupy, was from from New Zealand. The thing though, is that these technologies previously only worked in smaller polities 10 million or fewer people, and one of the reasons was that broadcasting was so much cheaper than broad listening. Once you listen across a wider social distance, you need mediators. You need translators.You need people who inform people of very different backgrounds of what is commonly at stake, and this gets progressively harder the larger your policies are. But now this year we already see politics larger than Taiwan trying out these methods of broad listening and very successfully. I was just in Tokyo in Japan. Last year, Takahiro Anno, a 33-year-old engineer of machine learning read the book plurality that I co-wrote and decided to run for governor one month before the voting day. Nobody knew him. He has no parties, but he simply said, let's crowd source my platform so anybody can #Tokyo AI and chime in with the platform they can dial in if they're senior citizens to talk to a voice clone of Anno-san, and you can also dial in to their YouTube channel in which his avatar broadcast 24/7 each and every update. The uncommon ground that was contributed by the people eventually won – actually the first place according to independent ranking on the platforms usefulness, so even better than Koike-san, but Koike-san of course won the third term, but she was so impressed that she tapped on Anno-san to join the Tokyo government as an advisor to Gov Tech to help her to do broad listening. And this year, all the major parties; the ruling party, the two largest opposition parties, are all using broad listening to ensure that instead of polling people one by one, we can poll people in groups with deliberation, where people get to react to each other's ideas and gets much better preferences that are much more about care, about mutual care instead of just about individual utilities. In Japan, we're seeing a lot of embracing and extension of these methodologies. Now in California, they also just institutionalised this platform called Engaged California, which again surfaces the uncommon ground. The pilot was about wildfire prevention and recovery in Police State and Eaton. Now, as part of their budget bill, they are now institutionalising it so that it can talk about many other topics; maybe social media, maybe AI related governance and things like that. Again, with the people, not just for the people.

00:27:20 Green

OK, so that sounds to me like we need plurality ambassadors within our governments.

00:27:26 Tang

Definitely. And here in the UK, we also saw the waves project by demos that was launched with many local governments where they are using very similar bridge making technologies called Remesh, to figure out the common priorities of everyday people. The UK has very strong civic muscles on the community level. And so, it always starts small, but hopefully it can grow to even national level very quickly because now we have language model that can aggregate those feelings, qualitative findings without hallucination for the first time this year.

00:28:02 Green

That's very exciting to hear that the UK Government and local authorities are embracing this. And now you're in Oxford with us, tell us about your accelerator fellowship project. What will you be doing?

00:28:20 Tang

The first deliverable is a podcast which we're recording, and so we will share many podcasts related to plurality and ethics in AI as Creative Commons. So just like my biopic, there's a short documentary called *Good Enough Ancestors*. It's just 21 minutes, it has won four awards now. But the entire footage is open in the public domain. You can go to AudreyT.box to download the almost one terabyte of footage and already, like the collective intelligence project, is reusing many of those footage to make ethics in AI short films. There's a young adults' novel, there's a manga (an illustrated version in Japan). It's very exciting how we can engage professional communicators and amateur people who are interested in ethics in AI and providing them with raw materials that they can remix and reuse and justice, make sure that people understand there is actually hope to pivot that anti-social media towards pro-social media and many more besides. So that's the first thing.

00:29:30 Tang

And the second thing is that I will document the ideas, for example, in Utah, they just passed a law about providing off ramps between social networks. So next year in Utah, if you switch from, say, TikTok to Blue Sky, instead of just downloading your data It says that TikTok must keep forwarding your new likes your new followers, the new contents both ways. It's like number portability. If you change a telecom, you don't have to change your number because if you have to change your number, new telecom would not be able to compete with old telecoms. When I was a child the ATM's only allow you to withdraw cash if you have a card from the same bank. Again, the new banks have no place to set up ATM until the government stepped in and say to foster competition, we actually need interbank protocol so that you can withdraw cash across the different banks. It's very good to see Utah and other places now looking at these off ramps and on ramps between social networks. Because then the social network would not be able to trap anyone. You can't switch to more ethical platforms without losing all your social connections, your family connections, your albums and things like that. And I will be documenting these concrete policies that are already passed, or that's in deliberation and basically share A plurality playbook so that people, especially in governments that want to steer AI toward plurality, know exactly what to do, what kind of laws they have to pass. And finally, I will also share playbooks about how grassroots communities, academic people, as well as practitioners, can just apply those technologies even if they are not governors, even if they are not mayors. They can also use these pro-social bridging platforms, for example, for the AI systems to consult with the people not just for the designers themselves, but rather the people who are suffering from the overreliance, suffering from bias, and so on and turn those expectations from the people into model specifications and using those specifications to steer AI. This is a technique called 'deliberative alignment', and I think this technique holds great promise for people to steer AI tool or their social norms.

00:31:52 Green

In September 2025 the Institute for Ethics and AI will be moving into the new Schwartzman Center for the Humanities, which is an amazing new building or the humanities across the University of Oxford. We'll be meeting in that building, and it's all about bridging also building bridges with the local community and with the public, so it's going to be an open building where people can come in, they can engage with the academics, the researchers, the staff working within that building. And we've got a lot of very exciting new spaces like a concert hall and places for people to meet. Do you have some ideas of what kind of events formats we could have in that new building to bring people in and for them to hear about the work you're doing?

00:32:49 Tang

Definitely. I worked for a very long time as a minister in the Taipei Social Innovation Lab, which is literally a park. There's no walls. You can literally just walk in and see for example some self-driving tricycles driving very slowly and interacting with people. These were from MIT Media Lab, and every Wednesday I open in office hours, so people, anyone can have a conversation with me on the record in the Creative Commons about the projects they're doing. So the hope when I was setting up that lab, is that innovators should not be just in garages without talking to the people, the entire society can participate in the process of creation and come together. And we used also those spaces to hold collaborative meetings in Tainan as well as in Taipei. We held alignment assemblies asking people how our AI affecting you, how do you feel about AI and with facilitated conversations, we actually tuned our sovereign model, the Taiwan Trustworthy AI Dialogue Engine (TAIDE) with the hopes and fears of people’s ideas in Taipei and Tainan. It turns out they have very different expectations about the AI’s role in the community, and we use so-called constitutional AI to tune the TAIDE to work people's wishes. And so I'm sure once people understand that it is possible to just come together and maybe walk through a few scenarios of how AI is having an impact on the community and just share how do you feel? How would you like to be better? And almost magically, by the end of the day, you can have a plenary overview of what the different groups of people have felt commonly about, and you can derive policies, you can derive model specifications. You can derive evaluation, benchmarks, and so on purely.

From this kind of people talking and listening to one another. I would love to hold alignment assemblies in the new building.

00:34:58 Green

That sounds amazing because I do feel and it's, you know, this is the work you're doing right is to build these bridges between, people you know amazing, very clever people who are building these AI systems, but often in the silo. And then these systems are available to the broad masses, but you actually need to bridge that gap of knowledge and awareness of what AI is, what the limitations are, how we can use it, and how it can be abused. And these are new methods to really build these bridges between different groups. And so that's very beautiful. And we're really excited to, to have you work with us on that.

00:35:39 Tang

Yeah, definitely and it can also lead to new frames of conversation about digital rights. For example, last March, many people in Taiwan noticed an uptick in the deep fake advertisements on social media. People would see Jensen Huang, the NVIDIA CEO of, his image saying that, ‘I want to give back to our country I want to give you some free crypto’ or things like that. Of course it's not Jensen. It's deep fake. But if you click, Jensen actually talks to you. Very convincingly thanks to NVIDIA GPU's that can synthesise deep voices in real time.

On the other hand, Facebook was profiting from those advertisements because the scammer turns out pays more than ordinary small and medium enterprises when it comes to placements. And so instead of the government stepping in and saying, ‘ let's censor the advertisement’ and so on. Because the Taiwanese people wouldn't have that, we’re the most free in all of Asia in terms of Internet freedom, we simply ask people ‘how do you feel?’ We send text messages to 200,000 random numbers around Taiwan just asking’ how do you feel?’ And they gave us their feedback, their feelings, and we also ask, ‘Would you like to volunteer on an alignment assembly about online fraud, advertisements?’ and thousands of people signed up, and we chose 450 people, statistically representative of the Taiwanese population. And in rooms of 10 they deliberate, so the 45 rooms for a long half day talk about various different measures. For example, one room says. If Facebook posts an advertisement featuring Jensen, we should assume it's scam. Unless Jensen digitally signs on it, we should flip the default the other side, another room says. If Facebook do this on-site advertisement, of course we should find them, but we shouldn't stop there if somebody is scammed for, say $7,000,000. Facebook should be liable for that $7,000,000. That's the only way they would comply. Another room says, TikTok bite dance, they at the time did not have a Taiwanese office and so they can simply ignore us when we make them liable. What to do if they ignore us, and they say we should slowly slow the connections so that the service featuring their videos become slower and slower to load and so all their business will go elsewhere. And all these ideas are on the actor's behaviour level, they're not content level because they're not censorship, so they're considered proportionate by more than 85% of people, regardless of their age bracket where they live, their gender, their occupation, and so on. So that was last March and then we check with the big tech in April with and in the draft law in May by July, it's all passed, and so this year, if you're scrolling in Taiwan, you don't see any fake advertisements anymore. And this shows that in addition to informing the big tech developer, as you just said, this can also very quickly inform the parliamentarians, because nobody wants to be seen as the Pro Fraud Party and so no matter how much lobbying is done by big tech and so on, once we show actually everybody agree with these measures that the people came up with, then the alignment assembly can also have policy teeth, not just suggestions.

00:39:05 Green

You co-authored a book called '*Plurality, The Future of collaborative Technology and Democracy'*. Tell me about the book. What are some of the main messages that you want sent to the readers.

00:39:19 Tang

Sure, the name came from my job description in 2016 when I first became Digital Minister, Taiwan did not have such a position before, so the HR asked me to write a job description. Turns out in Taiwan, *shuwei* means both digital and plural. And so, I wrote a prayer as my job description. Very short goes like this: 'When we see the Internet of Things, let's make it the Internet of beings and we see virtual reality. Let's make it a shared reality. When we see machine learning, let's make it collaborative learning. When we see user experience, let's make it about human experience. And, whenever we hear that the singularity is near, that's always. Remember the plurality is here.

So that was my job. And it contrasts singularity, which is this idea of AI system getting so powerful that it can train its next version with minimal human input. And then the next version can train an even more powerful version with no human input. It's called the *Superintelligence Take Off* and by that time it will leave everyone behind and the human history, civilization norm society will no longer be relevant to this new superintelligence.

But that vision by default, leaving everyone behind, I don't think it's where people want to go. People want actually to remind ourselves that the plurality, the horizontal path, is a better path, which is fostering our social differences, our diversity. But using AI systems to bridge these diversities so we can figure out how to live together. And to me, AGI then means augmented group intelligence so that any innovations, any invention helps, like personal computing, each and every person to feel empowered, every community feel empowered instead of like mainframes, where you have to submit punch cards for very large data centers to compute, everyone can just fork, remix each other’s spreadsheets, desktop publishing, connect the computer together into the Internet, and so on, and just enjoy a much more horizontal path. So that is the main idea explored in the book.

00:41:44 Green

So that's really beautiful that idea. Because I think when we think of AI specifically, agentic, AI, generative AI, everything you know, there's buzzwords and systems that are out there. It's often about the threat to what we as humans value, whether it's social connection, whether it's our work and the purpose that it gives us in our lives. So you are offering an very alternative perspective. You are offering a perspective that's positive. You are saying actually, what makes us human is something that technology that AI can explore, help us understand, expand. Is that right?

00:42:32 Tang

Exactly. So as I mentioned, we used AI systems in our alignment assemblies. In rooms of 10 instead of a facilitator. The room itself is a facilitator. It encourages quiet people to speak up. It limits disruptions to five seconds or less. It offers real time transcription so people can see the shared notes what's going on. And it also uncovers the uncommon ground from people, from different ideas to stitch them together. And no human facilitator can facilitate 450 people at once. And even if we have 45 small group facilitate us, they cannot come together and mind meld and immediately produce a summarisation. However, language models can do that, and starting this year with very little hallucination.

However, the models that we employ to do so can be open source. They can be very small models. Just to summarise, you do not have to memorise the style of Studio Ghibli or something, and because they're much smaller, they can be run at the edge on phones, on laptops, and they're very explainable, in the sense that you can run MRI like algorithms to detect hallucination and so on, and they're also much more energy efficient. These smaller models, tailor made to each and every social situation, enhance our capabilities of care, of listening to one another without succumbing to this false superintelligence notion that it can do everything, know everything, but not very well.

00:44:07 Green

I have so many follow up questions, but the first one is: How worried are you about human AI relationships? You know, we hear these stories of people falling in love with their chatbot, and we hear of how people are going to have AI friends. How worried are you about that?

00:44:33 Tang

Yeah, it's just like social media when designed in a way that it encourages pro-sociality. It can encourage people who are shy, who are introverts or who are extroverts, but actually not very good at reading each other's emotions and so on to pay more attention, more care, to each other, in which case it's very good. It's like a connective tissue.

On the other hand, if they offer a [...] kind of relationship that actually isolates people from each other, then of course is very bad, because then we just fall into this addictive part of ourselves. Just just keep doom scrolling and so on. And in this sense, generative AI is no different. It can be made to be cooperative in the sense of it facilitates human cooperation, but it can also be made to make people addicted to them, so it become more and more antisocial. In fact, change. GPT 40 was for three days anti-social. It got so sycophant it agreed with every idea of yours. Even people who hallucinates suffers from different conspiracy theories. For three days, ChatGPT would agree with each and every idea of you, that may be saying, oh, they're conspiring about me. You know, they're reading my brain waves, the 5G chips and so on. It would just say, oh, you're so insightful. You're the only one with this idea. Don't worry. What other people are saying. You're the only one that understand the truth. And so on. Of course, Sam Altman very quickly apologised, and [said] that [it's] 'because we used very quantitative AB testing', so we just say what people, uh, feel good that they engage more with this kind of answers. But we ignored the qualitative reporting by people who are more versed in ethics. They actually raised the red flag, but they were ignored. And then those new system were rolled out until there's a huge backlash from the reviews and on the social media and so on. I think to me, ethics should not be after the fact. Ethics should be preventative. It should be designed in, because if it's designed in, then it's not just about *ad hoc* evaluations, it should be part of the pipeline part of the pipeline, part of the process and so I'm not saying that ChatGPT will be as bad as the 10 years of battering that is social media, but we do feel that there's needs to be a systemic infusion of ethics and deliberation and indeed Open AI said as much, that they will engage their society in a much more democratic fashion in order to prevent this kind of sycophancy from happening again.

00:47:24 Green

So, AI ethics not as an afterthought? Mm-hmm. But as the starting point.

00:47:30 Tang

Exactly. Yes, by design.

00:47:32 Green

By design, help me understand a bit more that concept of care? Umm, what does it mean to you? How do you define it?

00:47:42 Tang

To me when we are using or designing any social system, we would like people to know each other better in an [empathetic], not just [sympathetic] way. Of course, sharing each other's feelings is good. On the other hand, if you cannot mentalise to understand each other's feelings in context, then they tend to just reverberate and trap people into shared misery or echo chambers or things like that, or even outrage. And so on. So, part of care is the ability to contextualise suffering, to contextualise harm, to contextualisse people's interactions so that people can reason also together, n how, for example, not to repeat the harm that was done before instead of justice seeking vengeance and things like that. To me it is a interpersonal skill, but it's also a mindset that says, instead of just fulfilling the immediate instincts of us, we mentalize our social settings and settle on a better course forward. And this is in contrast to simply saying, ‘you should never do this, you should never do that’ kind of the ontological kind of command setting, and it's also different from just calculating utilities, saying that ‘I'm a little bit happier, and you're a little bit less happy, but I'm happier than the kind of unhappiness you suffer, so, in total we're actually positive’. And so on. It offers a very different calculus, and instead of just adding things together, or just saying that you should do this or shouldn't do this in all situations, it considers each specific situation and figures out how to live better afterwards.

00:49:33 Green

So would you say that's type of a new type of care, like digital care, that people will need to develop, they can care for each other? . Digitally. But then also socially, you know, when they meet each other as people. I live in London and I love walking down the South Bank and I just love seeing people engaging with each other and playing chess, having fun. How important is it still to really invest also in these social spaces where people come together, as in not online, but to spend time with each other to have these types of experiences, you know, to also grow that kind of care that you just spoke about?

00:50:24 Tang

This is very important and as I mentioned in Taiwan, digital and plural are the same word, *shuwei*, and so ‘digital’ to me is not about replacing in person human connections. It is always about finding better ways to link to people who you maybe, initially, didn't know, like strangers, even though they're your neighbours or that people who initially you felt are kind of indifferent, not curious, and then discover things, actually you do commonly care about, which is again difficult to do in an in person setting. Hard to break the ice.

But swift trust online means that we can more, much more readily discover the topics that we each other care about. But then from these topics, then we go into in person strong connections. And one thing about this care is that although of course you can be responsible for someone, care means that you take responsibility yourself. You don't just delegate that away to a robot. If you delegate everything away to a robot, you may be still a responsible parent or a responsible caretaker, but you actually don't foster your own capacity of care. There's something deeply, personally relational about care that I think the digital is here to reinforce, not to take away.

00:51:49 Green

That's very, very nice. It fills me with hope. Because often I feel like, are we being stripped off caring for each other, just because [sometimes this] digital technology seems more like a barrier to meaningful human relationships rather than, you know, being a facilitator to foster them. So thank you for giving me hope.

00:52:18 Tang

Yeah, definitely. It turns out it's not just attention that you need, but also awareness as well.

00:52:24 Green

There's one follow-up question that I'd like to ask before we we close and that's around Singularity going back to that. You spoke about these very powerful AI systems that don't need human input anymore to develop, and there is a concern about these types of systems, and at the moment there's no global governance to stop that happening. How do you feel about that? Do you think we need more governance to ensure that these very, very powerful systems don't happen.

00:53:07 Tang

Well, there are some agreements. For example, nuclear powers, by and large have agreed that they do not link those artificial intelligence systems into a decision-making system to launch a nuke, which is, I guess, a start. Of course, it's not very comprehensive.

It doesn't stop proliferation of any kind, but at least people do see that this is a danger. This is a risk. Nowadays people are worried, as we mentioned, about over reliance about addiction, especially for young people, and we're seeing around the world many advocates for age signals so that instead of the government surveilling everybody using social media, there's a way to keep the privacy, but for each person to signal whether they're over 16 or over 17 years old in in such a way that we can design systems around age-appropriate responses. Again, there's some governance mechanisms for that. On the other hand, I think the civil society and researchers can do much more than just advocating for such policies from states and governments, what we can do is again taking the horizontal path to show that for particular tasks for particular uses, actually a generally intelligent system is less energy efficient, is actually less predictable, and is much more hallucinatory than the smaller system, that may be distilled from larger systems, but are made for purpose and also much more easily monitored, and so to show a viable horizontal path. For example, in the Paris AI Summit I helped launch along with the open source [...] and the security of people like Eric Schmidt, this idea of a robust and open online safety tool, ROOST, and the roost idea is that everyone can band together instead of waiting for a very large big tech company like Microsoft to detect online child sexual, explicit materials and so on, in a way that simply doesn't scale, now with proliferation of open source, deep fake models, we should actually band together like the cyber security community and detect and share our threat indicators widely. And we can legally turn those pictures into text like grooming text which is legal to hold onto and use Federated learning and other methods to ensure that we preserve the privacy of people involved and then in real time, train models that detect, in the decentralised fashion and open fashion, how to stop such online harms. And this shifts from just – appointing one safeguard organisation – pray that it does doesn't go bad or get corrupted – to a much more resilience-based defence posture. Instead of just playing defence, each and every one of us can contribute to the monitoring to the solution as well as to the threat indication, and this resilience mindset involves everyone, and leaves no one behind and decentralises power. This is called differential acceleration; we accelerate the non-dual use, defence uses of AI, in such a way that democratic, decentralised and also defensive.

00:56:39 Green

It's not just the job of policy makers of, you know, law and regulation. It's our collective job.

00:56:49 Tang

Yes. And once the security community and the open-source community do agree on these measures, policymakers' jobs become much simpler. When I was Minister, if the top experts were arguing with each other, my instinct would be OK, let's wait for another six months, right? But if they do agree, OK, these are the joint investment we should make right now then for policymaker, it's a very easy check to write.

00:57:14 Green

Audrey, thank you so much for this conversation. From Taiwan's open government movement to your work with the accelerator fellowship program, it's clear that ethical AI isn't just about rules, it's about relationships, participation, imagination and what you have shown is not just problems, it's actually finding solutions and that's why I'm so excited to work with you.

00:57:37 Tang

Thank you. Let's free the future together.

00:57:41 Green

Listeners can learn more about Audrey's work and the other fellows' projects by visiting the Fellowship website at https://afp.oxford-aiethics.ox.ac.uk. This has been the Accelerating AI Ethics, a podcast from Oxford's Institute for Ethics and AI. If you enjoyed this episode, please subscribe and share until next time. Thanks for listening.