Interview with Jerry Sheehan, Director for Science, Technology and Innovation, OECD

OECD Supports Hiroshima AI Process to Achieve Better Social Well-Being

By *Japan SPOTLIGHT*

International organizations are now striving to promote AI governance, as AI technologies demonstrate an unprecedented scope and scale of global impact on societies. Through good governance of this common challenge, we may find a way to better global governance in the post-Ukraine crisis world.

In the following interview, OECD Director for Science, Technology and Innovation Jerry Sheehan clarifies for us the OECD's pioneering work in this area that may help guide us to be a better world through better governance of AI.

(Interviewed on March 28, 2024)

Introduction of OECD Activity in AI Governance

JS: I would like to highlight the OECD's activity in this area, as international discussions on AI are very important, and the OECD is certainly a frontrunner on this issue. Could you give us a brief overview on the OECD's work on AI, such as the AI Policy Observatory and AI Incidents Monitor?

Jerry Sheehan

Sheehan: Our work on artificial intelligence began as early as 2016, and with a good push from Japan. That let us get started on trying to develop the OECD Recommendation on Artificial Intelligence, also known as the OECD AI Principles, and a report called "Artificial Intelligence in Society". Both were published in 2019. We've had now several years of very effective work on AI, and we like to say we've blazed the trail on policy making.

That report laid out a broad set of implications of AI for society and for our economies, and in particular, the development of the OECD recommendation which established a set of principles. In fact, the OECD Recommendation on Artificial Intelligence was the first international intergovernmental standards on AI, intended to be flexible, future-proof and values-based. It also promotes the development of a more human-centric AI.

Of those five principles, the first is about pursuing using AI for beneficial outcomes for people and the planet – so already the recommendation has been anticipating concerns about climate change and other environmental issues. The second principle deems human-centric values and fairness as being essential. Thirdly, it calls for transparency and explainability, so that we can understand how AI algorithms work, how they've come to their conclusions, and on which data they were trained. The fourth principle promotes the robustness, security, and safety of AI systems so that we can have confidence the systems themselves will operate in ways that are anticipated. The fifth principle is about accountability and ensuring that there is a set of responsibilities for AI systems and system developers, and those using it.

We like to think that those Principles have been very influential in the broader policy environment. In fact, we've already seen elements of them taken up in the European Union's AI Act. And we've seen other elements of our work that have been captured in areas like the National Institutes of Standards in the United States. They've also been very influential in helping move forward the Hiroshima AI Process that was initiated by Japan under its presidency of the G7 just last year. We've had the OECD AI Recommendation in place now for five years and we're in the process of reviewing and revising the recommendation. Part of that is informed by the work that we've been doing to better understand how countries are approaching AI and AI governance issues.

As you mentioned, we have the OECD AI Policy Observatory, which is a platform open to the public, in which we collect information from countries about their AI-related policies, their AI national strategies, and other specific initiatives. As of today, we have information on more than 1,000 different strategies and initiatives across 70 countries and jurisdictions, so extending well

beyond the OECD's 38 member countries and the EU. We also use that as a platform for gathering real time data related to AI. We have information available there on investments in research and in venture capital, as well as around AI skills and the kind of job requirements that are being posted related to AI. In addition, we have some tools to use to visualize the data. So that helps us track how the AI ecosystem and the policy environment are changing.

We also know, of course, that there are risks and concerns associated with AI. We did some good work in support of the Hiroshima AI Process in 2023, looking across the G7 countries and to understand what these issues and concerns are and to help develop an evidence base. We launched in November 2023 at the Paris Peace Forum something called the AI Incidents Monitor. It uses AI and other tools to help monitor reliable news sources for incidents where AI has, for example, propagated misinformation or disinformation or where it has caused safety harm or concerns or led to issues of so-called deep fakes and images. It has allowed us to track these events in close to real-time and to help us get a better sense as to how frequent these kinds of events are, what types of events are occurring, and where. This is another piece of our tooling to support policy making.

We've also developed something called the OECD Framework for the Classification of AI Systems, which helps people think about different types of AI. We know that there are technologies like natural language processing that are used in a variety of applications. Generative AI is maybe the most recent example. We have image-processing applications that are often used, for example, in the medical domain. So this is a framework to help us think about the different kinds of risks and opportunities associated with different kinds of AI and to manage those appropriately.

We've also created something called the OECD AI Network of Experts, a multidisciplinary group of more than 400 experts. They come from government, business, academia, and civil society, and from OECD member countries and beyond. This provides a way of getting early signals of new developments in AI, and of bringing the expertise of these groups to help us identify potential solutions.

These data and tools complement what we can do at the OECD through our more typical committees and working parties which have government representation, and help us keep abreast of what's important and happening in the AI space.

JS: OECD principles are certainly very influential in developed nations, but they do seem to have started affecting non-member countries' policies as well. And to my knowledge, OECD principles do not have any enforcement mechanism, but since this is a good model, many countries are following the principles. Is that correct? And if so, does it perhaps amount to a kind of peer review mechanism?

Sheehan: That's a very good question. For the OECD AI recommendation, as with other OECD recommendations, we establish standards that we think can contribute to good governance, but we don't have formal enforcement mechanisms. There are two ways to think about it, as you were suggesting: the notion of peer review and providing transparency into how countries are using our legal instruments are important elements of our work.

Peer review is, for example, done through the AI observatory where countries are providing information about their strategies and about policy approaches that they are taking. We also systematically review the recommendations. This typically happens on a five-year cycle; a lot has happened with AI in the last five years, even just over the last 18 months with generative AI. The review process gives us another opportunity to work with our member countries to examine how they are implementing the OECD AI Principles, where they're having challenges with doing so, and where they think new developments and updates may be necessary.

At the same time, the principles have become more widely adopted than just within the OECD countries, and we are encouraging other countries to take them up and adopt them as well. In fact, the OECD AI Principles provided the foundation for the set of AI principles that have been adopted by the G20 countries.

And here at the OECD, we serve as the secretariat for the Global Partnership on AI (GPAI), which was also established five years ago by the G7 countries – again with support from Japan. GPAI was echoed in the most recent G7 outcomes and provides a platform for the G7 countries to work together to advance the adoption and implementation of the OECD AI Principles. GPAI now has 29 countries as members – 23 of them are OECD member countries, and it works with six other countries and hope that number will grow over time. GPAI member countries include India, which currently chairs the group, Argentina, Brazil, Senegal, Serbia, and Singapore. That gives us another ability to extend the influence of the OECD AI Principles.

I think the other element that's important to remember is that countries and regions like the EU can take the Principles and embed them in their own national or multinational legislation, which then allows them to establish enforcement and monitoring mechanisms. We've seen this already as part of our revision to the OECD AI Recommendation, for example. We made a first step in adopting and updating the definition of what an AI system is, taking into account some of the developments around generative AI. That definition itself has now been incorporated into the EU's AI Act. As we look at that act and at other legislation that is developing in other countries, we can see traces of the OECD AI Recommendation in those, but now with an enforcement mechanism.

And then I think what we can try to do as well, as many countries and regions are working toward implementing governance mechanisms, legislation, and regulation around AI, is to work to develop common terminologies, common frameworks, and common approaches. That way, even if these are done in individual regions, the regulations should be more consistent with each other. We like to call them interoperable, so that they can work to an extent across borders and across boundaries.

JS: One of the most impressive points about your

remarks is that your principles are rather flexible in response to the various needs of AI. And that's very good to know because the rules are fixed and cannot easily be changed. That could have some negative impact on the progress of AI or the progress of technology. However, sometimes I have difficulty in understanding the difference between AI and other IT technologies. What would make a distinction between the two?

Sheehan: AI is a type of IT technology and it's getting quite a bit of attention now because of generative AI, which I think brought AI into the public space. We need to recognize that AI has been embedded in a number of different IT tools for quite some time. AI systems have the ability to take data that can be provided to it or sensed and generated, and to make predictions or sometimes take some action, usually with some human oversight to implement an outcome.

We look at AI systems, which span the realm from using natural language processing to image processing to technologies like neural networks that generate computational technology outcomes. Compared to other IT systems and so forth, it has a higher potential for autonomy and self-improvement as well. For example, we use training sets of data to help AI systems to be able to look at images and detect which images might show a sign of malignant cancer versus those that are benign. The systems themselves can learn over time as they're presented with new data.

That's one of the key areas of differences between AI and other IT systems, like what the one we're using right now for this interview, or like the IT systems that we use for doing word processing or spreadsheet analysis, which are much more constrained in their abilities and they don't have this ability to learn over time. A key element around AI is its ability to improve over time and to make inferences. And now, as we're seeing as well, what makes AI more distinctive is the ability to generate novel content – but of course, always based on the content that it has been trained on in these so-called large language models.

Maximizing Positive Aspects of AI & Minimizing Negative Aspects

JS: Moving to the question of positive and negative aspects of AI, particularly generative AI, the underlying principles should be well balanced between promoting the positive and restricting the negative. Do you think raising productivity by AI could mitigate the challenge posed by, for example, depopulation in Japan?

Sheehan: As you say, our common refrain is that we want to capture the benefits and mitigate or minimize some of the risks and challenges. Productivity gains are certainly one area where we see opportunities for AI. Equally, we see opportunities as we see potential improvements in health care and in the way we educate and train our students. And of course, another area where we see great opportunity is around scientific discovery and accelerating our ability to make discoveries and then translate them into better health, better education, and higher productivity in the workplace. In some cases, we see this quite clearly – we're seeing generative AI, for example, being used as a tool to help software engineers and those who write software code. They're using AI and it's helping them decrease the amount of time they need on a specific task by more than 50%.

There's a huge productivity gain. And we've been trying to get a more systematic understanding of how AI is going to affect work, productivity, and skills through something we call our AI WIPS Program on work, innovation, productivity, and skills. We've surveyed 7,000 different employers and workers on the impacts and done a number of case studies. What we're finding so far is that there are productivity increases; typically with AI, we find that it's more a process of reorganizing work to embed AI rather than displacing work and reducing the number of workers.

But that also means we need to ensure that workers have the skills and retraining they need to use these new AI-powered tools. What we're seeing is that a lot of workers and employees, although they have concerns so far, have been positive about the impact of AI on their own performance and their working conditions. At the same time, yes, there's concern about longer-term job losses and we do need to take steps to make sure that we are building trust in these AI systems through training and consultation with our workers.

In the face of slowing population growth or declining populations, AI can help us maintain, or even increase in some ways, our levels of productivity given that the workforce gets the training and that we restructure how people interact with IT systems to get most of that benefit.

JS: My observation of AI is that my productivity has been raised because I can collect more data and more information without much difficulty. This is empowering but it doesn't necessarily lead to a reduction of employment, because individuals will be empowered but they don't necessarily have to worry about their jobs being lost or replaced by AI. Also, somebody mentioned that there might be some concerns about the digital divide between those who can use AI very well and those who cannot. This was a concern during the IT revolution, but that may not be the case with AI because there doesn't seem to be such a high barrier to using AI – you can just click on ChatGPT and get a load of information. Would you concur with this?

Sheehan: The issue of the digital divide comes up in two ways: one is at the individual level and the other is at the firm level. Because not all jobs necessarily deal with manipulating and processing information the way that you and I do, some jobs could be considered more at risk than others in terms of job loss. But I do think that in many cases, as with other elements of IT, we see that

the technology is being used to improve the work that is done by individual workers and workplaces. At the individual level, we need to ensure that people have the right skillsets to adopt and use AI. So, there is a training component to address the digital divide.

There are differences in the ability of firms to adopt and make use of AI, so we need to be cautious about digital divides across the economy, similar to what we see at the individual level. We see that the firms that are best able to adopt and make use of AI for productivity tend to be large firms or young startups that tend to have more highly skilled employees and more complementary assets, such as better digital infrastructure. So there is a need to ensure that regional divides driven by differences in broadband deployment don't further accentuate overall productivity divides, and there are opportunities for workforce training to ensure that companies can better benefit and take advantage of the opportunities that AI offers.

JS: Another negative aspect could be disinformation. We would need to overhaul the legal system or intellectual property rights. If driverless cars are more common, we would need more regulations for driving with regard to liability for accidents and so on. In that sense, would comprehensive regulatory reform be necessary everywhere in the world? If so, will we need more efforts for regulatory harmonization in international venues and does the OECD have such concerns?

Sheehan: It's an important question and as you know, there are several international efforts going on right now to help shape the global governance of AI. Certainly, our efforts here at the OECD, but also G7 countries have been quite active in this space, and similarly the G20. There was the UK AI Safety Summit in 2023, and two more AI safety summits are planned for 2024 and 2025 in South Korea and France. And of course, the United Nations has also been involved in these issues of AI governance. I think there is a question of how we ensure that these international efforts, which are complemented by national efforts, are moving in the same direction. The questions of harmonization are challenging, so we want to try to ensure that we are all moving in the same direction and that we are developing regulatory approaches that are based on some common understanding. I like to think that the common ground would be the OECD AI Principles and that they can promote interoperability between different initiatives.

Within the G7 meeting last year at Hiroshima, there was a clear intention to develop a common course of action among G7 countries regarding generative AI and produce a comprehensive policy framework that has both guiding principles and a code of conduct for the developers of advanced AI systems. There is an ambition now to develop codes of conduct that would apply to users and implementers of AI systems and all stakeholders in the AI ecosystem. This code of conduct is something that we are trying to frame across countries to implement the shared principles.

We also see in the EU some political agreements on its AI Act which will affect all EU countries. As I mentioned, this AI Act is a good example of drawing on the OECD definition of an AI system and classifications in the OECD AI Principles tools like the OECD Framework for the Classification of AI Systems, which we have put in place to support risk management across different types of AI systems. Harmonization and coming up with a single uniform approach across all countries can be challenging, but this notion of improving interoperability between the regulatory regimes in different countries and regions is going to be increasingly important.

We are already putting some thought into what we can do to help companies and others who might be affected by the code of conduct; how we can help them in a consistent way – regardless of which country they are from or operating in – to demonstrate their compliance with this code of conduct. I think these efforts can span across national boundaries and are the kind of work that international organizations can lead.

We need to do this in a truly global fashion, and that is why we are working with the Global Partnership on AI with other countries well beyond the OECD member countries. We have had some interactions very recently with nations in the African Union and we are trying to understand how to best develop and adopt AI in their region. Being inclusive and ensuring that we engage the full set of stakeholders from researchers and developers of AI systems to those who deploy them in different sectors of the economy, and to those involved in education and training, is going to be incredibly important.

International Cooperation Needed to Address the Issue Properly

JS: Are you already working with other international organizations on this issue?

Sheehan: We have been working with many other international organizations as we become aware of them and as I mentioned, we have provided quite a bit of support to the G7 through Japan's Hiroshima AI Process last year, and we are continuing to work with Italy through their 2024 G7 presidency. We are engaging with the G20, which adopted the OECD AI principles, and we are continuing to work with them on the next steps – especially as our revised recommendation will be ready in 2024. We have been supporting the UN efforts as they are now working on the Global Digital Compact and have identified several ways in which we can contribute work to help support that effort within the UN, as well as with the AI Safety Summits that have been in place. We have participated in both of those at a high level from the OECD, and we are identifying other ongoing work that we can contribute. We certainly recognize that AI is an issue that needs global engagement and global cooperation, and we are ready to collaborate with other international partners as well as provide our inputs to these international fora.

JS: AI is a global challenge, like the global environment, so might we need to reach a new

consensus on international collaboration in the post-Ukraine war world with AI?

Sheehan: I think that might be an ambition to have a global consensus around it considering that there are countries that have different values and different approaches that they want to take. We've been working in the OECD with countries that also have a similar values-based approach to AI development and then engaging beyond that set of countries on particular issues of mutual interest. But let's hope over time that, even though we have different approaches, we can reach that consensus.

JS: That does seem to be a growing argument on restricting AI, but as we mentioned at the beginning, we would need some balance between restriction and promotion because AI is certainly an important tool for raising productivity innovation, but what do you think about this trend?

Sheehan: I think we need a balanced approach that both promotes AI development for all of the reasons and opportunities we see and mitigates the risks associated. Boosting productivity, improving our education systems, improving our health care, accelerating research and so forth – all of those are reasons why we want to accelerate AI development and diffusion. We need to promote AI in an informed way by understanding what the challenges are both in terms of digital divides and the potential to accentuate inequalities within societies and countries. We have growing concerns about misinformation which generative AI can contribute to, and there are fundamental safety risks. There are concerns about AI systems that have significant control over important infrastructures or important decisions. Thus we need to put in place guardrails to help ensure that AI development moves and stays on track, but at the same time, we need to be building the road itself that's going to take AI and help develop it.

I like to go back to the AI principles on this as well because, in addition to the high-level principles that we've established about risk, transparency, and safety, we've made several recommendations to countries in implementing these. The first of those is that we need to continue to invest in R&D for AI. We know there is potential in AI technologies and we have been investing in R&D in some countries going back to the 1950s and 1960s, and now we are seeing a quantum leap in the capabilities of AI. But there's still more to discover and we need to build the ecosystem that will help us understand what's coming next.

We need to make sure that the research community is working with developers and educators as well as with civil society to understand what the concerns are and to anticipate them. We need to put in place a good policy environment that is both enabling and protective. We need to foster international cooperation, and be attentive to the concerns about labor market effects to ensure that we are using AI to improve opportunities for workers, and in fact, provide them with the opportunity to adopt it. We are trying to come up with a balanced approach to promote AI while ensuring that we understand the concerns, and we are taking informed steps to address them.

OECD's Future Work to Promote Well-Being Created by AI

JS: My last question is about the OECD's future work. Well-being is a concept created by the OECD, but what would be your future work in AI-related issues to promote well-being?

Sheehan: That's a great question because in the end, we want AI to contribute to overall societal well-being. The directorate that I lead here (Directorate for Science, Technology and Innovation) has been looking at AI and trying to understand its implications for a long time. Right now, our immediate future work is the completion of the revision of the OECD AI Recommendations and Principles. The principles have held up very well at the high level over the last five years, but of course, with the rise of generative AI, we see certain concerns that have been accentuated, such as misinformation and disinformation and heightened concerns about job displacement.

So, we are hoping now that the revised recommendation will be ready to be adopted at the Ministerial Council Meeting of the OECD in May 2024 which is being chaired by Japan. We will continue to work to support the next steps on the Hiroshima AI Process, which was launched last year but is continuing within the G7 this year too.

We recognize as well that AI is being addressed in virtually every other part of our organization. Our Directorate for Employment, Labour and Social Affairs is looking at the effects of AI on labor markets, social protection and health care, and the Public Governance Directorate is looking at how public sectors can use AI better. Our Directorate for Education and Skills is looking at how we can use AI to improve overall skills development training and the approach that we take to education and learning. The OECD is taking a more multidisciplinary approach to understanding AI and its impacts in the future, all guided by the notion that we want to maximize benefits and mitigate risks but unleash the potential in all of those areas and more across the spectrum.

We are trying to do more internally to coordinate our efforts and to understand how AI is affecting society. We are taking the "whole of government" and a "whole of OECD" approach, as well as evidence-based and data-based approaches to issues around AI. Some of the tools that we are putting in place will collect more information on job opportunities, on AI incidents, and so on, so that we can make informed recommendations to governments; certainly, to our member countries with the aspiration to spread them beyond. The wellness aspect is certainly part of our endeavor, as we look at the ambition of for the OECD to advance "Better Policies for Better Lives". We want to do that around our AI work as well. JS

Written with the cooperation of Joel Challender who is a translator, interpreter, researcher and writer specializing in Japanese disaster preparedness.