Journal & Proceedings of the Royal Society of New South Wales, vol. 157, part 1, 2024, pp. 119–122. ISSN 0035-9173/24/010119-04

Doing AI well: the Responsible AI network

Stela Solar

Director, National AI Centre, CSIRO[1]

stela.solar@data61.csiro.au

I really love the expressions that Sally was using, especially the unravelling of the data and patterns and causality and so forth. I landed in technology by complete accident. I was going to be a film composer and I loved creativity and self-expression and how I could help enable others to do the same. Finishing my University degree, I had to start adulting, getting a job, and I landed in a technology start-up by a complete accident. I learned on-the-job, completing certifications, many courses and then I bolstered that with a Master's. What I found was that technology was so incredibly creative for even my own interest in self-expression: during my Master's study I developed an emotion-sensing dress that would change colour and shape based on how you were feeling. It would augment your own expression. The same with an interactive sleep cocoon, that would be connected to your biological processes: change shape, use vibrations, use binaural beats, so that you get the maximum sleep over the night. I've been fascinated by how technology can actually augment ourselves, how we work, how we express ourselves and so on.

Somehow I got into helping industry succeed with technology, leading the National AI Centre, hosted by CSIRO. Believe it or not, we don't have any researchers at the National AI Centre — there are lots of AI researchers at CSIRO, but at the National AI Centre we are working with commercial organisations every day to understand how they're using AI, the challenges they're encountering, and helping them implement AI responsibly. Don't get too connected to that word "responsibly" — some people call it "ethics," some people call it "trustworthiness," some people call it "safety, diversity, inclusion." Industry wants to do AI well but it's quite a challenge figuring out how exactly to do that today.

The industry context that we're hearing at the National AI Centre is that the AI narrative is very polarising. Right now, it's either all incredibly high-risk or there's great optimism that it's going to solve everything. The reality is that, depending on the use case, it's somewhere in between. Some use cases are low-risk and in fact have been around for a very long time. In Sydney, for instance, some of the infrastructure and transport solutions out in the world today have been leveraging machine-learning, data science, AI for 40 years. It's actually how we as humans in our day-to-day lives are able to continue making informed decisions in highly complex environments. What we often hear from business is that, with AI becoming so polarising, some of these basic use cases that are making sense, patterns and predictions, are also sometimes deemed to be seen as high-risk, but actually are kind of very basic and non-risk.

In addition to organisations using AI to navigate complexity, there are also some

1 This is an edited transcript of the talk [Ed.]

major global challenges that we're tackling. Some of you might have seen CSIRO's seven megatrends that are shaping our next decades of life and work.[2] Some of these challenges are tremendous: no matter how many of our brains and hands we connect, we could not tackle them. There are not physically enough health professionals to provide quality care to people who need it. Or climate change and adapting to climate change. This is such a complex ecosystem dynamic that we're needing the greatest of our technologies to tackle that.

In the commercial sector AI is seen as this great solution to help tackle and unravel complexity so we can continue making informed and meaningful decisions. I want to share two key examples in the health space: this is a particular area where I think AI is going to add so much value because it can augment our ability to make sense of the world, find patterns, and take actions on them.

The two examples: one is Hive Health in virtual environments. It's implemented at Royal Perth Hospital. In essence, there is a pod of four medical professionals who are monitoring 200 patients remotely. They're gathering vitals and health data so that they can see who is needing medical intervention. It's helping the doctors be more effective: rather than doing walk-by checking on the patient one by one, it's helping the medical professionals go to where they need to be. They're leveraging this AI Hive solution to augment their ability to make an impact by being where they need to be.

Another interesting case is the work by Dr Helen Fraser in leveraging AI for breast-cancer mammography. Dr Fraser has built machine-learning models to help spot anomalies in mammograms. Rather than thinking about replacing the medical professionals who might otherwise be looking through these mammograms, she has optimised this model to look at the most basic use cases, and either rule out the existence of an anomaly, or detect an obvious anomaly. It frees up time for the medical professional to focus more on the highly detailed complex cases. It's a real way of thinking about how machine learning and AI tools can augment the professionals for that impact.

Quite often when we talk about AI, very quickly the conversation turns to bias. AI hasn't brought new bias to us: bias has always been in the world. But because AI is built on data, it can often propagate biases unless it's designed responsibly. There are two examples that I want to use which have helped us see the flip side, where AI can actually help tackle some of the biases that we may not even see. One example is EY (Ernst & Young), who have a loan-approval solution that they provide to banks. It's one of their services. There's some automated risk scoring that might be presented to the financial service organisation they're working with. But what they found was they used an AI model called FAIR-learn and they found that there was bias in the data they were using for loan approvals. In fact there was a 7% disadvantage that women had during the loan approval process versus men. And this system had been operating for a long time without AI, but now with AI was able to find that bias. The same toolkit was used to reverse some of that bias, so it has moved from 7% to 0.3%.

2 https://www.csiro.au/en/news/all/news/2022/july/seven-megatrends-that-will-shape-the-next-20-years [Ed.]

Another fascinating example is a solution from Sapia.ai, an Australian company that has a chatbot for early-stage interviewing. The headline in the *Financial Review* (May 17, 2023) said, "AI more likely to hire women than humans are," but I think the most interesting data point is when you dig a little bit deeper into the study, conducted in collaboration with a university in the UK. They found that when women were told that they were being interviewed by an AI bot and assessed by an AI bot, 37% more women applied, which is starting to suggest that we have biases around us. There are members of our community who would feel more fairly treated potentially by AI systems, if those systems are designed responsibly and fairly.

I haven't even spoken about generative AI yet because AI has been around us since the '50s. Generative AI has somehow made it seem like suddenly AI is a new thing, and the last year has completely changed everything. What has changed is the ease with which every single person can engage with AI systems. Suddenly many more people are using AI systems to augment what they're doing: to get creative ideas, to help them draft a first email. In fact our signals are showing that in the workplace 30 to 40% of employees are using generative AI. 68% are not telling anyone about it. That's really fascinating to think about.

Put yourself in the shoes of a business leader: you know your people are using it — there's some productivity signal in there that your people are finding more effective ways of doing work. But if you don't know about it, that brings exposure to your organisation, because you don't know what people are sharing, you don't know which services they're using. That is why right now the first thing that we suggest to commercial organisations is that they must implement a generative AI policy because, no matter what your perspective on it is, it's happening. I think one of the highest risks for any team, any organisation, is to have hidden dynamics and hidden use. Even if it if something is not exactly according to strategy, they would rather know about it than have it hidden.

The generative AI use cases that we often hear about are things like customising sales emails and personalising advertising and so forth.[3] But I want to share two that are a little bit out of the ordinary: one is generative AI for cybersecurity. Currently the average Australian is getting more than 250 cybersecurity attacks a year. That's huge, and the challenges are increasing for organisations as well. Our teams at Data61 leveraged generative AI to create cybersecurity honeypots — files that seem like they have very valuable and confidential data, or they might have personal information, or they might have credit cards or some IP. The bad actors, who are attacking technology systems, want to get to this data; they want to monetise it or sell it or exploit it. The Data61 teams at CSIRO built a generative AI tool that became very good at creating honeypots but completely fake ones to distract the bad actors and draw them to the fake "precious" data, and so keep them away from the real precious data.

Another case in design: some of you might have seen that NASA is using generative AI to create parts for their satellites. So you

3 Indeed, this transcript has been edited by ChatGPT 3.5 with the following instructions: "Edit the transcript of a speech. Eliminate the uh and um words. Use British spelling. All sentences end with a period — capitalise the initial letter of the sentence." [Ed.]

define: "this is the dimension of the part that I need, don't put anything here, this is where hands need to go, don't put anything here, this is where the sensor pack needs to go, this is where the attachment is." It needs to be very specific. And, for the rest, "generative AI, draw a strong structure as light as possible using this material." Generative AI colours it in or draws it in. Interestingly, they're finding that these structures are more resilient and stronger than they've been designing before. I encourage you to look at these offline in your own time, because the designs look organic. It's not something that has come out an angle-ruler engineering approach: there is a real finesse. That's intriguing.

It's not only NASA that's doing this — even Shell is using generative AI to design the layout of its wind farms. It says, "Hey, generative AI, this is my terrain, this is the altitude, this is the weather, this is the weight of my wind turbines. Tell me the layout options." This is one way that people are using generative AI to help tackle the complexity and decision-making in highly tangled data environments.

Just to wrap up: if we're relying on such technology so much, then we need to ensure that they're trusted, that they're accurate, that they're safe. Much of the experience with generative AI has been in the consumer-facing products space. They're trained on very broad, uncontrolled data sets. I think there's much more opportunity for generative AI on controlled organisational data to help people find what they need.

It does need to be trusted. Today, more than 74% of organisations around the world are not even checking their data for quality or bias. More than 65% are not checking data drift or model drift. We have a need to actually develop and level up our practice of doing AI well. That's what we focus on at the National AI Centre and why we develop the Responsible AI network:[4] it was to share some of this best practice of how to do AI well, so that, when we do choose to use AI to augment our processes or decisions, we can do so in a in a more trusted and responsible way. Thank you.

4 https://www.csiro.au/en/work-with-us/industries/technology/national-ai-centre/responsible-ai-network [Ed.]