Source: *The Straits Times,* pA22
Date: 5 July 2019

How to train AI to be fair to humans

Singapore's launch of an AI governance framework gives it an opportunity for global thought leadership in regulating AI. But issues of fairness and explainability in AI are more complex than commonly realised.

Thiparat Chotibut and Yap Jia Qing

For The Straits Times

Minister-in-charge of the Smart Nation Initiative Vivian Balakrishnan spoke at this year's Budget debate about the Singapore Government's plan to "double down" on artificial intelligence (AI).

As companies and government agencies start to take guidance from the Model AI Governance Framework that Singapore launched at the World Economic Forum earlier this year, it is important to note that progress on AI governance is still far from complete, as acknowledged by the framers themselves.

Two concepts advocated by the framework are "explainability" and "fairness".

While they appear commonsensical in insisting that AI rules should be capable of explanation to humans and appear fair, in fact, the technical and ethical notions of these concepts remain highly debatable.

These challenges are not only issues that Singapore will have to face as it pursues a nationwide roll-out of AI, but they also provide opportunities for Singapore to demonstrate thought leadership on a global level, if managed well.

ISSUES FOR A MULTICULTURAL SOCIETY

One issue in which Singapore can shape global norms in AI is racial bias. Algorithmic bias is a major challenge. Sensitive social issues such as racism can result, if premature technology adoption precedes careful examination.

For instance, a photo-tagging algorithm misclassified images of African descendants as "gorilla".

New Zealand's automatic passport photo verification system incorrectly rejected an Asian man's application because his eyes appeared "closed".

A recent study from the Georgia Institute of Technology revealed that pedestrian detection models in autonomous vehicles trained on standard data sets are more likely to miss detecting people with darker skin.

In a multiracial and multicultural society like Singapore, the social fabric can be put at risk if unwarranted systemic discrimination is discovered to be in-built, even inadvertently, into AI systems.

Although there are numerous toolkits for evaluating and mitigating problems with AI fairness, these examples raise suspicion that perhaps the current problem in AI fairness is not due to the lack of technical and quantitative solutions.

Rather, the difficulty is in defining what fairness considerations need to be taken into account in a particular application and community.

A research paper accompanying IBM's open-source fairness toolkit noted that "(t)here is a need for a large number and variety of fairness metrics in the toolkit because there is no one best metric relevant for all contexts. The metrics must be chosen carefully, based on subject matter expertise and worldview".

Human judgment is still required as a bastion of fairness for a particular society and this goes into decisions regarding the curation of data sets used to train AI models, for model tuning and for algorithm audits.

Last year, MIT Media lab, through its Moral Machine experiment, crowdsourced the opinions of citizens globally on their choices between two unavoidable outcomes in a hypothetical road accident, with one being whether an autonomous vehicle should spare the young or the elder.

The results suggest that the decision-making process of autonomous vehicles, and AI systems in general, should be culturally informed.

Differences between individualistic cultures and collectivistic cultures were observed. In the latter, which emphasise the respect that is due to older members of the community, people showed a weaker preference for sparing younger characters.

As the research also acknowledged, while the ethical preferences of the public should not necessarily be the primary arbiter of policy, the willingness of the public to procure and tolerate AI in society will depend on the palatability of the ethical rules adopted.

THE QUESTION OF FAIRNESS

If fairness is so important a consideration in developing ethical rules for AI, we need to then consider what fairness might mean in Singapore society.

A starting point would be to look at the factors laid out in the Singapore Constitution where discrimination is not permitted – religion, race, descent or place of birth. Discrimination against Singaporeans based solely on any of these factors in law, in the appointment of public officials, or in the administration of some laws, could be a potential breach of the Singapore Constitution.

An informal understanding of fairness – that similar people should be treated similarly – has emerged in the computer science community.

Coincidentally, this is similar to the legal notion of equality that the Singapore courts have adopted in interpreting the Singapore Constitution with regard to equality before the law.

All the law requires is that like persons in like circumstances should be treated alike – the law may discriminate between classes in certain cases, but no individual within a particular class should be singled out for discriminatory treatment.

This points to the reality that some form of discrimination might be needed (or tolerated) in the governance of society, and thus AI as well. The challenge lies in determining what distinguishing factors are permissible for drawing the lines between these classes, between which discrimination is allowed to happen.

BALANCING EXPLAINABILITY WITH ACCURACY

Getting to a precise notion of explainability has been a subject of debate among philosophers, linguists and computer scientists alike, but its essence is the belief that a system, processes or rules being set up, must be capable of being explained to a person. This includes being able to answer why an AI model predicts or decides the way it does. Then there is the separate question of what makes an explanation "good" enough.

Various approaches to explainability have been attempted by researchers, including creating simpler models used solely as a proxy for what a more complex model is doing, or having AI models output a sentence that captures the main reason why it decided the way it did. Some other practitioners simply advocate for the use of more transparent models instead, often at the expense of prediction accuracy.

The challenge in making AI more explainable has been in balancing between completeness and interpretability: that is, between having a complete explanation that represents the intricacies of the model, and having an explanation that can be interpreted and understood in human terms.

The most complete explanations are often difficult for humans to interpret, while more interpretable explanations are often oversimplified.

Just like how a simplified drawing of the human anatomy in a primary school textbook, though easy to understand, might not always be helpful in helping us understand what is happening when we are struck with a certain rare ailment.

The only difference with AI is that with millions of moving parts, or what researchers call parameters, in a model across different dimensions, even AI researchers can struggle.

If the objective of explainability is to build trust in an AI system, it might not be ethical to present a simplified abstraction of the whole system, if the assumptions behind the simplification are not clear to users.

WAY FORWARD FOR SINGAPORE

The launch of the laudable Model AI Governance Framework at the World Economic Forum, and the consequent global attention it brought to Singapore in the regulation of AI, has been a great start in demonstrating Singapore's potential to be a thought leader in the space.

Also encouraging are the developments in the setting up of the Advisory Council on the Ethical Use of AI and Data, the AI Singapore initiative and research centres at the law faculties of local universities.

Actually fulfilling the ambitious ideals laid out in the framework would require a multi-disciplinary approach, engaging the technologists involved on the ground in building and tuning AI models on a day-to-day basis – who are, and would be, the ones making the actual decisions that matter.

The open consultation on the Model AI Governance Framework has been a good first step.

Lawyers, social scientists and philosophers also have much to offer in helping determine, if not the answers, then the questions that need to be asked in the design of AI models relevant to the Singapore context.

stopinion@sph.com.sg

• Thiparat Chotibut is a theoretical and computational physicist at Chulalongkorn University in Thailand. Yap Jia Qing is a research assistant on regulating artificial intelligence at the National University of Singapore's Faculty of Law. They are co-founders of the Emerging Technologies Policy Forum, a Singapore-based cross-disciplinary think-tank advocating for technically informed policy and legal discourse.