RSiS S. RAJARATNAM SCHOOL OF INTERNATIONAL STUDIES
Nanyang Technological University, Singapore

RSIS COMMENTARY

www.rsis.edu.sg

No. 128 – 2 July 2019

*Singapore Defence Technology Summit*

# AI Ethics 2.0:
# From Principles to Action

*By Danit Gal*

## SYNOPSIS

*Trying to catch up with the fast moving and increasingly pervasive development and use of AI, nations want to establish ground rules to ensure the technology benefits humanity. This is no easy task, further complicated by the mismatch between abstract AI ethics principles and existing technical capabilities and human practices.*

## COMMENTARY

IN THE past couple of years, discussions on AI ethics became the norm, with a variety of actors putting forth over 40 sets of fairly identical principles. These principles include: accountability, controllability, diversity, explainability, fairness, human-centricity, transparency, safety, security, and sustainability.

While this creates a shared language assisting countries in addressing similar concerns, the local interpretation of these principles can differ widely, often leading to a deep sense of confusion. With the wide proliferation of such AI ethics principles, practitioners are getting closer to agreeing on what they should be in theory, but not on how to make them work in practice.

### Problem of Effective Implementation

Principles like the ones recently published by the OECD are illustrative. They combine many existing works on AI ethics principles into another fairly generic guideline. This level of abstraction makes it appealing enough to create an important international consensus.

It is also, however, vague enough to allow local actors to interpret the principles as they see fit within their own social and cultural contexts. This diversity of interpretation is essential in ensuring that benefits brought to humanity by using AI are inclusive. The problem is, therefore, one of effective implementation.

The problem of effective implementation is the test bed of these principles, which can often be detached from technical capabilities and human practices. Can these AI ethics principles be codified into technical and human practices? In theory, yes. The abovementioned principles benefit us. They are intended to keep us safe and help us all benefit from the use of AI.

In practice, however, they face real-life conflict of interests such as corporate profitability, individual and collective biases and inequality, low general levels of technical literacy, the sanctification of progress, and the desire for constant convenience.

## Moving from Principles to Action

The good news is that this problem is already being partially solved. Individuals and institutions working in the AI ethics field are moving from principles to action. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems led to the creation of a series of technical standards on AI ethics. The IEEE is the world's largest association of technical professionals.

The Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) community brings together researchers and practitioners developing tangible solutions. The Machine Intelligence Research Institute (MIRI), the Center for Human-Compatible AI (CHAI), and safety teams at DeepMind and OpenAI work to develop safe and robust AI. Institutions like AI NOW, Data & Society, and various academic centers are getting to the heart of socio-technical problems and how they already impact users.

The bad news, however, is that while this work contributes immensely to the beneficial development and use of AI, it is still only the beginning. We need wider geographical participation and action to put AI ethics principles into practice and create inclusive benefits. Until we are able to achieve that, any benefits called for in AI ethics principles run the risk of staying as an idealistic vision for a speculative future.

## Coming to Terms with the Present

While AI might still look futuristic, its early stage applications are already as widespread as they are pervasive. To most, AI is invisible. Users cannot really see it or interact with it, and they often do not understand how it works or affects them and their actions. And yet, most regulations and AI ethics principles look towards the theoretical future and thus fail to address the implementation programme.

Due to their soft governance nature, most AI ethics principles do not offer tangible solutions. More alarmingly, many government regulators remains unwilling to offer tangible solutions due to fears that overregulation will 'stifle innovation'.

This creates a false dichotomy where ethical and well-regulated developments ‘sacrifice’ speed or innovation to ensure benefit. In reality, not making this ‘sacrifice’ leads to development that is prone to structural errors, stalls in achieving market viability, and mostly just serves its developers.

Additions to the over 40 existing sets of AI ethics principles are a positive and welcome development if they represent new concerns and population groups. But things will only change when local governments interpret and implement them in local regulations and more institutions develop technical tools and methods to put them into practice. Tangible solutions are within reach.

## The Small Country Advantage

Smaller countries have an edge in solving this problem. As importers of technology from larger countries, smaller ones often find themselves relying on tools not developed with their social and technical needs in mind. This entails an adjustment and adaptation period for users and the technology itself.

All governments must, therefore, invest in creating regulatory and technical sandboxes to ensure the adjustment and adaptation period goes as smoothly as possible and comes to positive conclusions. But small governments can do it faster and more efficiently. To that end, they should do two things:

The first is to institute agile regulatory mechanisms that develop with and support the nation’s beneficial use of AI. The second is to invest in creating well-informed and resourced actors that put local AI ethics principles interpretations into practice.

Investment in a competitive future should be about beneficial development, not just a rapid one. If not, our future will see us spending years trying to identify and fix the mistakes we have made in the name of careless progress, and that’s the best case scenario. In short: put well-considered theory into thoughtful local regulatory and technical practice, because 40+ sets of AI ethics principles will not work unless you do.

*Danit Gal is founder of the TechFlows Group technology geopolitics consultancy, and creator of the Collective Futures Network for young experts. She is a researcher working on AI ethics, safety and security. She contributed this to RSIS Commentary in cooperation with RSIS’ Military Transformations Programme.*