# ARTIFICIAL INTELLIGENCE FOR SOCIAL GOOD

Keio University

APRU

# ARTIFICIAL INTELLIGENCE FOR SOCIAL GOOD

# Contents

# Philosophical point of view for social implementation

## Institutional and technological design development through use of case based discussion

## How to expand the capacity of AI to build better society

# Foreword

**By APRU**

The dual character of artificial intelligence technology, its promise for social good, and its threat to human society, is now a familiar theme. The authors of this report note that "the challenge is how to balance the reduction of human rights abuses while not suffocating the beneficial uses". Offering a solution, they go on to say that "the realization of social good by AI is effective only when the government adequately sets rules for appropriate use of data".[1]

These observations go to the core of the challenge before all societies. Whose interests do governments mainly represent? Are they accountable in real ways to their citizens or are they more aligned to the interests of high-tech monopolies? As with all technologies, we face the questions of ownership and of their use for concentrating political power and wealth rather than ensuring the benefits are shared with those most in need of them.

The current COVID-19 crisis has shown that governments need to move decisively towards the public interest. We confront crises within a new economic order of information technology that "claims human experience as free raw material for hidden commercial practices"[2]. The multidisciplinary studies in this report provide the knowledge and perspectives of researchers from Singapore, Hong Kong, Korea, Thailand, India, and Australia that combine the local understanding with the international outlook that is essential if policymakers are to respond with appropriate regulation (and taxation) to ensure technology companies with a global reach are enabled to contribute to the common good. The insights in these chapters underpin the report's recommendations on developing an enabling environment and a governance framework.

This is the third in a series of projects[3] exploring the impact of AI on societies in the Asia-Pacific region which offers research-based recommendations to policymakers. It is intended that the reports support the work towards achieving the UN Agenda 2030 for Sustainable Development and its goals.

Subsequent work might usefully look at the ways that social movements can assist formal regulatory processes in shaping AI policies in societies marked by inequalities of wealth, income and political participation, and a biosphere at risk of collapse.

This project is a partnership between APRU, UN ESCAP and Google. International circumstances permitting, we will work together to hold a policy forum later in 2020 or early 2021 to share these findings with policymakers and public officials from around the region.

I thank our partners for their support and Professor Jiro Kokuryo, Vice President of Keio University, Tokyo, along with members of the Project Advisory Group for their leadership of this initiative.

**Christopher Tremewan**
*Secretary General*
*Association of Pacific Rim Universities*

1. Introduction p.4

2. Zuboff, S. (2019). *The Age of Surveillance Capitalism*. See 'Definition' in opening pages

3. *AI for Everyone* (2018) led by Keio University; *The Transformation of Work in the Asia-Pacific* (2019) led by
   The Hong Kong University of Science and Technology. **https://apru.org/resources/**

**By UN ESCAP**

In 2015, governments agreed on the 2030 Sustainable Development Agenda to "ensure peace and prosperity, and forge partnerships with people and planet at the core". In this global agenda, science, technology, and innovation were identified both as a goal in itself and as a means of supporting the achievement of other sustainable development goals.

Artificial intelligence (AI) offers a myriad of technological solutions to today's problems, including responding to COVID-19, enabling better delivery of public services[1], and supporting smart innovations for the environment. However, the wave of optimism surrounding the transformative potential of AI has been tempered by concerns regarding possible negative impacts, such as unequal capabilities to design and use this technology, privacy concerns, and bias in AI.

The world must ensure that AI-based technologies are used for the good of our societies and their sustainable development. Public policies play a critical role in promoting AI for social good. Governments can regulate AI developments and applications so that they contribute to meeting our aspirations of a sustainable future. Governments, in particular, are

encouraged to invest in promoting AI solutions and skills that bring greater social good and help us "build back better" as we recover from the impacts of the COVID-19 pandemic.

While much has already been written about AI and a world of possibilities and limitations, this report is based on realities and experiences from Asia and the Pacific, and provides various perspectives on what AI for social good may look like in this region. More importantly, the report offers suggestions from the research community on how policymakers can encourage, use, and regulate AI for social good.

I look forward to more research collaborations with ARTNET on STI Policy Network[2] – a regional research and training network supporting policy research to leverage science, technology, and innovation as powerful engines for sustainable development in Asia Pacific.

**Mia Mikic**
*Director*
*Trade, Investment and Innovation Division*
*Economic and Social Commission for Asia and the Pacific*

---

1. *Artificial Intelligence in the Delivery of Public Services* (UN ESCAP, 2019).
    **https://www.unescap.org/publications/artificial-intelligence-delivery-public-services**
2. **https://artnet.unescap.org/sti**

It has been a great pleasure for Keio University to take the academic lead in such an important initiative as the UN/ESCAP-APRU-Google project "AI for Social Good". We are extremely pleased that the joint efforts of government, academia, and industry have generated a set of academically robust policy recommendations.

In our efforts to overcome COVID-19 with the help of information technology (IT), we are reminded of the importance of having a firm philosophy on the use of data. For example, we have seen first-hand the effectiveness of IT-based "contact tracing" in controlling the spread of the disease. At the same time, we are uncertain about the technology and its implications on privacy. There are noticeably different views on this topic concerning data and privacy, with cultural differences playing a major role. Some cultures are happy to actively share data, while others place greater emphasis and value on protecting privacy. At the same time, although all cultures recognize the value of sharing data, they are seemingly split on whether the data should belong to society or the individual. The design of technologies and institutions vary depending on such fundamental philosophies behind the governance of information. We do not, however, want the world to be split along this divide, as this leads to the fragmentation of data and everyone loses out. In order to benefit from the great technologies that we possess, the world must come together.

Since Keio University was founded by Yukichi Fukuzawa in the middle of the 19th century, we have been a pioneer in introducing Western

thought to Asia. During his life, Fukuzawa advocated the introduction of Western culture to Japan and placed great emphasis on relationships between people for the creation of a modern civil society. Today, this would encompass the idea of harmonious coexistence between people and technology. From such a heritage, we are cognizant of our renewed mission to bridge differences and create a new civilization that makes full use of data while honoring the dignity of each and every person. Of course, this is easier said than done. In reality, we face competition among nations and businesses who all have interests in controlling, monopolizing, and/or profiting from data. We should also be alert to the possibility that technologies can actually widen rather than close the inequality gap between the haves and have-nots.

With this in mind, academia should pledge to stay loyal only to evidence and logic. Through such self-discipline, we can provide open forums to orchestrate collaboration among various stakeholders to work together for the good of humanity. This is a worthwhile endeavor, as we are certain that artificial intelligence has the power to solve many issues, including epidemics, and will help us to achieve the Sustainable Development Goals proposed by the United Nations.

**Akira Haseyama**
*President*
*Keio University*

# Introduction

# Artificial Intelligence for Social Good

Yoshiaki Fukami and Jiro Kokuryo
Keio University

## 1. Harnessing AI to Achieve the United Nations Sustainable Development Goals

We live in a complex world in which various factors affecting human wellness are interconnected and cannot be analyzed by simple models. For example, solutions to the challenges of pandemics require understanding of not just biology and/or medicine but of social activities, as well as the psychology of people who spread groundless or even malicious rumors on social media.

Expectations are high that artificial intelligence (AI) can help develop solutions to many issues facing the world by identifying patterns in the vast body of data that is now available through today's sensor networks. By enabling machines to identify and analyze patterns in data, we will be able to detect issues and causal relations in complex systems that were previously unknown. Such knowledge is essential in our efforts to overcome complex issues.

We should also be mindful that both wellness and these complex issues are embedded in local contexts that are diverse and depend on geographic and social backgrounds. While recognizing such diversity, it would be useful to have a meta-level understanding of how AI could be applied to accomplish our goals. An integrated and comprehensive vision, as well as its related policies, are needed to realize effective approaches for more people to enjoy the benefits of AI.

With this in mind, the United Nations (UN) has already begun to take a higher-level approach to solving social issues with AI. Set at the General Assembly (2015) and to be accomplished by 2030, the UN Sustainable Development Goals (SDGs) look to harness AI in support of inclusive and sustainable development while mitigating its risks. For example, SDGs look to:
• Provide people with access to data and information
• Support informed evidence-based decisions
• Eliminate inefficiencies in economic systems, as well as create new products and services to meet formerly unmet needs
• Provide data-driven diagnoses and prevent harmful events such as formerly unpredictable accidents
• Support city planning and development

This report understands AI for social good as being the use of AI to support SDG achievement by providing institutions and individuals with relevant data and analysis.

Table 1 is a non-exhaustive list of initiatives by the UN and other institutions to use AI in support of achieving SDGs. Supplemented with additional examples, the table mainly presents initiatives included in the UN Activities on Artificial Intelligence report by International Telecommunications Union (ITU, 2019). While the table presents projects that use AI for social good, it does not include initiatives that attempt to mitigate the risks of AI, such as to address bias or other ethical concerns.[1]

| SDG | Use of AI |
|---|---|
| 1 No Poverty | • Implementation of AI on the Global Risk Assessment Framework (GRAF) to understand future risk conditions to manage uncertainties and make data-driven decisions (ITU, 2019, p.54) |
| 2 Zero Hunger | • FAMEWS global platform: Real-time situational overview with maps and analytics of Fall Armyworm infestations (ITU, 2019, p.3)<br>• Sudden-onset Emergency Aerial Reconnaissance for Coordination of Humanitarian Intervention (SEARCH), and Rapid On-demand Analysis (RUDA) using drones and AI to greatly reduce the time required to understand the impact of a disaster (ITU, 2018, p.54) |
| 3 Good Health and Well-being | • Ask Marlo: An AI chatbot designed to provide sources for HIV-related queries in Indonesia (ITU, 2019, p.22)<br>• Timbre: a pulmonary tuberculosis screening by the sound of the cough (ITU, 2019, p.22) |
| 4 Quality Education | • AI to ensure equitable access to education globally: Provide hyper-personal education for students and access to learning content (UNESCO, 2019, p.12)<br>• Using AI and gamification to bridge language barriers for refugees: Machine learnt translation for lesser-resourced languages (UNESCO, 2019, p.11) |
| 5 Gender Equality | • Sis bot chat: 24/7 information online services to women facing domestic violence (United Nations Women, 2019) |

**Table 1: Notable initiatives using AI in support of achieving SDGs**
*(Created by Daum Kim)*

---

1. It should be noted that most projects supporting *Goal 5: Achieve gender equality and empower all women and girls* focus on removing gender bias. We only found one initiative using AI to empower women – a project that uses AI to fight against domestic violence.

| 6 | Clean Water and Sanitation | • Water-related ecosystem monitoring through the Google Earth Engine and the European Commission's Joint Research Centre to use computer vision and machine learning to identify water bodies in satellite image data and map reservoirs (ITU, 2019, p.32)<br>• Funding analysis and prediction platform using Microsoft's Azure Machine Learning Studio to capture global funding trends in the areas of environmental protection by donors and member states (ITU, 2019, p.32) |
|---|---|---|
| 7 | Affordable and Clean Energy | • Mitsubishi Hitachi Power Systems (MHPS) in the development of autonomous power plants: A real-time data monitoring action to reduce supply or increase generation and automated capability to manage power plants (Wood, 2019)<br>• Intelligent grid system to increase energy efficiency through AI (Microsoft & PwC, 2019, p.17) |
| 8 | Decent Work and Economic Growth | • Analysis of the impact on jobs and employment by investigating the rise and effect of reprogrammable industrial robots in developing countries, along with exploration of patent data in robotics and AI to understand the future impact of AI robots on work (ITU, 2019, p.9) |
| 9 | Industry, Innovation, and Infrastructure | • E-navigation: Exchange and analysis of marine information on board and ashore by electronic means for safety and security at sea (ITU, 2019, p. 13)<br>• Maritime Autonomous Surface Ships (MASS): Attempts to apply automated ships (ITU, 2019, p.13) |
| 10 | Reduced Inequalities | • Implementation of AI in a Displacement Tracking Matrix (DTM) to detect and contextualize data such as migration, urban and rural land classification, and drone imagery in displacement camps (ITU, 2019, p.16) |
| 11 | Sustainable Cities and Communities | • Risk Talk: An online community to exchange climate risk transfer solutions. AI builds a neural network by mapping the expertise of the users through interactions on the platform (ITU, 2019, p.37)<br>• United for Smart Sustainable Cities initiatives (U4SSC): A global platform for smart cities stakeholders which advocates public policies to encourage the use of ICT to facilitate smart sustainable cities transition  (ITU, 2019, p.29) |
| 12 | Responsible Consumption and Production | • AI-driven system and robotics to reduce food waste by predicting customer demand (Fearn, 2019)<br>• iSharkFin: Identification of shark species from shark fin shapes to help users without formal taxonomic training (ITU, 2019, p.3) |
| 13 | Climate Action | • Shipping digitalization and electronic interchange with ports (ITU, 2019, p.12)<br>• Cyber-consistent Adversarial Networks (CyberGans) to simulate what houses will look like after extreme weather events to allow individuals to make informed choices for their climate future (Snow, 2019; Schmidt et al., 2019) |

**(Cont.) Table 1: Notable initiatives using AI in support of achieving SDGs**
*(Created by Daum Kim)*

| 14 | Life Below Water | • Maritime Single Window (MSW) to electronically exchange maritime information via a single portal without duplication (ITU, 2019, p.12) |
|----|------------------|---------------------------------------------------------------------------------------------------------------------------------------|
| 15 | Life on Land | • DigitalGlobe's Geospatial Big Data platform (GBDX) using machine learning to analyze satellite imagery to predict human characteristics of a city and respond to health crises (ITU, 2018, p.50)<br>• Land governance and road detection through satellite "computer vision" (ITU, 2018, p.60) |
| 16 | Peace, Justice, and Strong Institutions | • International Monitoring System of Comprehensive Nuclear-Test-Ban Treaty Organization (ITU, 2019, p.1)<br>• Toolkit on digital technologies and mediation in armed conflict (ITU, 2019, p.27) |
| 17 | Partnerships | • The International Telecommunication Union (ITU) Focus Group on AI for Health (FG AI2H) (ITU, 2019, p.19)<br>• The AI for Good Global Summit: Identifying practical applications of AI towards SDGs (ITU, 2019, p.19)<br>• Social Media Data Scraper: AI on natural language processing helps to understand the thoughts of users (ITU, 2019, p.38) |

**(Cont.) Table 1: Notable initiatives using AI in support of achieving SDGs**
*(Created by Daum Kim)*

## 2. Report Objectives: Research-based Policy Suggestions

Having reviewed how AI can be applied to promote social good, we now turn to policies that adequately promote and control AI, so that they can be used for the good of society. This is important, as we believe our goals cannot be accomplished through a laissez-faire approach. An adequate governance system for the development, management, and use of AI is crucial in ensuring that the benefits of integrating and analyzing large quantities of data are maximized, while the potential risks are mitigated.

Following an agreement between APRU, UN ESCAP, and Google to share best practices and identify solutions to promote AI for social good in Asia-Pacific, the project AI for Social Good was launched in December 2018 at the Asia-Pacific AI for Social Good Summit in Bangkok. Each chapter of this report presents a unique research project (Table 2), as well as key conclusions and policy suggestions based on the findings. The projects were selected following a competitive process that sought research inputs to inform policy discussions in two broad areas:

1. Governance frameworks that can help address risks/challenges associated with AI, while maximizing the potential of the technology to be developed and used for good.

2. Enabling environment in which policymakers can promote the growth of an AI for Social Good ecosystem in their respective countries in terms of AI inputs (e.g., data, computing power, and AI expertise) and ensuring that the benefits of AI are shared widely across society.

Focusing on specific local contexts and with the objective of informing international policy debates on AI, the research reports offer a range of unique perspectives from across the Asia-Pacific region.

| Chapter | Title | Resaerch Member(s) | Affiliation |
|---|---|---|---|
| 1 | AI for Social Good: Buddhist Compassion as a Solution | Soraj Hongladarom | Chulalongkorn University, Thailand |
| 2 | Moralizing and Regulating Artificial Intelligence: Does Technology Uncertainty and Social Risk Tolerance Matter in Shaping Ethical Guidelines and Regulatory Frameworks | M. Jae Moon Iljoo Park | Yonsei University, Republic of Korea |
| 3 | Definition and Recognition of AI and its Influence on the Policy: Critical Review, Document Analysis and Learning from History | Kyoung Jun Lee | Kyung Hee University, Republic of Korea |
| 4 | Regulatory Interventions for Emerging Economies Governing the Use of Artificial Intelligence in Public Functions | Arindrajit Basu (Team leader) Elonnai Hickok Amber Sinha | Centre for Internet & Society, India |
| 5 | AI Technologies, Information Capacity, and Sustainable South World Trading | Mark Findlay | Singapore Management University |
| 6 | Governing Data-driven Innovation for Sustainability: Opportunities and Challenges of Regulatory Sandboxes for Smart Cities | Masaru Yarime | The Hong Kong University of Science and Technology |
| 7 | Including Women in AI-Enabled Smart Cities: Developing Gender-inclusive AI Policy and Practice in the Asia-Pacific Region | Caitlin Bentley | University of Sheffield, Australian National University |
| 8 | AI and the Future of Work: A Policy Framework for Transforming Job Disruption into Social Good for All | Wilson Wong | The Chinese University of Hong Kong |

**Table 2: List of project titles and their authors**

The AI for Social Good Project believes that objective, evidence-based, and logical academic analyses which are free from political and/or economic interests can play critical roles in the formation of sensible policies. At the same time, we are aware of the tendency of academics to stop at simply understanding the phenomena and not take a position in prescribing policies. Hence, we specifically asked the participants of this report to come up with short summaries of their findings, as well as suggested policy implications (see Appendix 1).

We also firmly believe in the effectiveness of a multi-disciplinary research approach for policy formation. To that end, the project organizers were careful to include both the technical and social sciences/humanities. We are extremely happy to report that all of the diverse teams, who shared a similar passion for taking a multi-disciplinary approach, were able to conduct fruitful discussions which led to even stronger projects.

# 3. Overview of the Recommendations

Based on discussions with the project members, this section presents the editors' own overview of the policy agenda, giving readers a general idea of the issues that need to be addressed.

## 3.1. Developing a governance framework

### 3.1.1. Ensuring equality and equity

In Chapter 1, Hongladarom makes an important suggestion in that policymakers should start by agreeing on the basic principles for the governance of data. That is, he discusses how altruism, as opposed to individualism, should be seen as the guiding principle to realize the benefits of data sharing. He also emphasizes its usefulness in correcting existing social and economic inequalities, which may expand with advances in technology. While this assertion may be controversial, it nevertheless addresses the fundamental question of whether data should belong to the individual or society, since we know that the value of data increases as they accumulate. This line of thought is also significant in that it reflects the communal traditions of Asian societies.

In Wong's discussion of AI's impact on employment (Chapter 8), he also calls for social security policies and a fair re-allocation of resources in the governance of AI. The editors' interpretation of such calls for social equity surrounding AI is that there may be strong scale advantages in AI (or data) economy that give unfair advantages to already powerful entities; and that policy intervention is necessary for fairness and to ensure the productive power of AI is able to materialize. Bentley's call (Chapter 7) for the inclusion of women as beneficiaries of AI is also along the same lines.

### 3.1.2. Managing risk to allow experimentation

All of the researchers recognize the potential for AI to both benefit and cause harm to society. The problem is, we will not know for sure what the positive and the negative impacts might be until we test them. It is therefore necessary to formulate a bold strategy to realize full potential of AI and manage the risks involved at the same time.

In Chapter 6, Yarime looks at the possibility of taking a "sandbox" approach to testing. In this way, experimental use of technology can be undertaken for proof of concept in a controlled environment, and the results can then be used to take the technologies outside the "box" to be implemented in societies at large. He also discusses the importance of preparing mechanisms for compensation, such as insurance, to mitigate damage done to individuals or institutions despite all necessary preventative measures having been taken. This function is crucial, not just to protect citizens but also to promote innovation.

Uncertainty and unpredictability are inherent characteristics of emerging technologies and cannot be eliminated completely. It is worth remembering that we should not sacrifice innovation through excessive safety precautions. If we want to benefit from technological advancements, we must be willing to take certain risks. As such, we should be thinking about "managing" risk rather than "avoiding" risk.

### 3.1.3. Multi-stakeholder governance and co-regulation

In Chapter 2, Moon and Park call upon the participation of different stakeholders representing industries, researchers, consumers, NGOs, international organizations, and policymakers in setting guidelines for the ethical use of AI. Most AI applications require cooperation of multiple organizations, particularly in the preparation of integrated datasets. For example, automobile driving data from a car manufacturer are only useful when combined with other data sources. The value of such data is further enhanced when combined with data from local and national governments that control infrastructure, such as traffic lights. Each of these actors have different objectives and, in the absence of adequate incentives, tend to tailor their systems to maximize the effectiveness of their own services without regard for the needs of others. Thus, not only do we need mechanisms to promote collaboration, governments should play a role in preparing them.

Although a natural temptation under such circumstances is to centralize control, we must also be aware of the dangers of a centralized approach both technically and societally. On the technical side, centralized databases are vulnerable to attacks and can result in large-scale data leaks once the system is breached. On the societal side, a monopoly over data gives excessive power to the institution that controls it, raising fears of a breach of human rights. A multi-stakeholder governance structure involving government, non-profit organizations, industry groups, and specialist groups should be established to provide oversight of the major players controlling the data. It is important that young policymakers and engineers participate in the discussion (Chapter 5). Given the rapid advances in technology, we must also develop and establish governance mechanisms that can evolve in a timely manner.

### 3.1.4. Providing accountability

Basu, Hickok, and Sinha (Chapter 4) identify accountability as one of five major areas where states should play a role. This is an extremely important point in light of the fact that AI can easily become a "black box" both technically and institutionally.

Accountability is a fundamental issue across various aspects of AI utilization, from the collection of data to the determination of evaluation functions in AI algorithms. As such, it is vital that we review and evaluate the process by which AI functions, as well as identify appropriate entities to manage the technology.

Accountability must be realized not only through legal systems, but also in the technical specifications of systems that ensure transparency of data management. Due to the pace of technological advancement, this is a challenge. Hence, governments need to assist in the development of a coordination mechanism that can cope with the progress in a timely manner.

## 3.2. Developing an enabling environment

### 3.2.1 Correctly understanding the technology

In Chapter 3, Lee cautions that, before discussing policies concerning AI, we should first have a proper understanding of the definition of AI. He points out the dangers of perceiving AI as simply machines that imitate and replace humans. Instead, he favors the perspective of the Organization for Economic Co-operation and Development (2019) that defines AI as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments" to form adequate expectations for the benefits of the technology.

An adequate definition of AI is therefore important, as it greatly influences the design of the governance

structure around the technology. Whether or not we recognize "intelligence" and "personality" (or at least legal personality as we recognize corporations as pseudo-personalities) in machines that seemingly have an intelligence of their own is becoming a serious topic of debate. If we are to adopt Lee's argument, then perhaps we should not.

### 3.2.2. Ensuring universal access to data

In Chapter 5, Findlay looks at how information asymmetries can create inequities for disadvantaged economies, and calls for systems to guarantee them access to data which enables them to negotiate fairly in international trade. This reminds us that AI cannot work on its own. In the application of AI, datasets, computing power, and expert analysts are all necessary to meet society's needs.

Naturally, the opportunities which computer networks create should not be underestimated. Recent advances in the reduction of communication costs, improvement of computing capabilities, and diffusion of sensing technology have facilitated the generation of big data that can then be analyzed by data scientists. Findlay's concern over inequity is especially important as there still remain many areas where access to essential data are limited and necessary data analyses are not possible. No matter how sophisticated the AI algorithm, it can only work effectively in an environment in which the dataset is properly generated and stored for analyses, there is the necessary computing power, and there is reliable and affordable access to expertise and the Internet.

It is worth remembering that network ubiquity does not exist yet either. There are still many people in Asia-Pacific that do not have access to reliable, affordable, and high-speed Internet. As such, governments should continue their efforts to provide everyone with Internet connectivity so that they have access to the data that empowers them.

### 3.2.3. Standardizing data models

Standardization of data formats is important in order to ensure universal access to data for a more equitable use of the technology. Not only does the differences in data models (formats) hinder data integration, a lack of standardization nullifies the power of ubiquitous Internet connectivity that enables us to gather data quickly and cheaply. In other words, aggregated data does not automatically mean big data suitable for AI analysis. Data must still be standardized to be collectively meaningful. In addition, data specifications (e.g., syntax and vocabulary) facilitate interoperability among distributed data resources and enable the generation of relevant big data. Furthermore, quality criteria enable data consumers to appropriately handle diversified data resources.

However, standardization is a complex issue, not because it is technically difficult but because it is a political process involving many different stakeholders, pursuing different goals. Therefore, a top down approach to forcefully impose a single set of standards will not work. That said, governments should still play a facilitator role, together with many non-governmental standardization initiatives, to prevent excessive proliferation of standards across every sector of society. Governments should also ensure interoperability among systems that of different standards.

### 3.2.4. Universal access to human resources for utilization of AI

Findlay also stresses the need for adequate assistance (e.g., technology, training, and domestic policy advice) to fully realize the benefits of AI. This is a reminder that AI systems require people to function. In other words, effective use of AI requires people to fine tune the algorithm and prepare the dataset to be fed into the system. It is also necessary for people to interpret the outcome and give it practical meaning. As the use of AI grows, so too does the demand for data scientists who can use the technology for social good.

However, as data scientists are fast becoming an expensive human resource only available to more developed economies and large corporations, the fewer number of them in less fortunate communities is limiting the opportunities to make use of AI.

When talking about human resources, it is important to recognize that not just software engineers and expert statisticians need to be trained. Senior executives and ordinary people also need to be aware of the benefits, risks, and mitigation measures surrounding AI, so that they are better informed and able to take advantage of the technology.

Another aspect is the need to educate engineers about the ethical, legal, and social implications (ELSI) of AI. As the power of AI grows, so too does its impact on ELSI. For the technology to be developed and used properly, governments need to ensure that technical experts are educated to be sensitive to the concerns of ordinary people concerning AI.

### 3.2.5. Removing the fear of using personal data

Another policy goal that the editors would like to propose is the removal of (perceived) risk associated with personal data disclosure. We believe that it is important to make available as much data as possible for the use of AI for social good. Of course, this is only achievable when people feel safe about disclosing their information.

There are two main reasons why citizens and consumers are currently holding back from offering their data for social good. First, they fear that data disclosure can lead to discrimination. This is especially true in socially sensitive areas. For example, when disclosure of infection to a disease leads to exposure to social stigma and criticism for non-compliance to social norms, people will be reluctant to cooperate with contact tracing. Second, certain consumers dislike the idea of having their data commercially exploited without their consent.[2] For example, the emergence of target marketing as the key revenue generator for online businesses has led to significant hostility towards the use of personal data.

To address this issue there are technical and institutional solutions available. On the technology side, various forms of anonymization, encryption, and distributed approaches in managing data have been proposed. Institutionally, various forms of regulations are in place to protect individuals from breach of privacy. For both types of solutions, government involvement seems essential in light of the incentives that exist, particularly in the private sector, to keep data secret for financial reasons. Not only should incentives be offered to make data public, but enforcement power must be used in the protection of privacy.

---

2. We should also be aware of people who are willing to give their information away for free, because they feel compelled or see a benefit in doing so.

# References

International Telecommunication Union. (2018). United Nations Activities on Artificial Intelligence (AI) 2018. https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2018-1-PDF-E.pdf

International Telecommunication Union. (2019). United Nations Activities on Artificial Intelligence (AI) 2019. https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2019-1-PDF-E.pdf

Microsoft & PwC. (2019). How AI can enable a sustainable future. https://www.pwc.co.uk/sustainability-climate-change/assets/pdf/how-ai-can-enable-a-sustainable-future.pdf

OECD 2019, "Recommendation of the Council on Artificial Intelligence", https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Schmidt, V., Luccioni, A., Mukkavilli, S.K., Sankaran, K., Bengio, Y. (2019). Visualising the Consequences of Climate Change Using Cycle-Consistent Adversarial Networks. https://arxiv.org/pdf/1905.03709.pdf

Snow, J. (2019). How artificial intelligence can tackle climate change. https://www.nationalgeographic.com/environment/2019/07/artificial-intelligence-climate-change/

United Nations Educational, Scientific, and Cultural Organization (2019). Artificial intelligence in education, compendium of promising initiatives: Mobile Learning Week 2019 https://unesdoc.unesco.org/ark:/48223/pf0000370307

United Nations General Assembly. (2015). Transforming our world: the 2030 Agenda for Sustainable Development. https://doi.org/10.1163/157180910X12665776638740

United Nations Women. (2019). Using AI in accessing justice for survivors of violence. https://www.unwomen.org/en/news/stories/2019/5/feature-using-ai-in-accessing-justice-for-survivors-of-violence

Wood, J. (2019). This is how AI is changing energy. https://spectra.mhi.com/this-is-how-ai-is-changing-energy

# Appendix 1

# Summaries of Papers and Policy Suggestions

## AI for Social Good: A Buddhist Compassion as a Solution

Soraj Hongladarom, Department of Philosophy, Faculty of Arts, Chulalongkorn University

### Abstract

In this paper, I argue that in order for AI to deliver social good, it must be ethical first. I employ the Buddhist notion of compassion (*karunā*) and argue that for anything to be ethical, it must exhibit the qualities that characterize compassion, namely the realization that everything is interdependent and the commitment to alleviating suffering in others. The seemingly incoherent notion that a thing (e.g., an AI machine or algorithm) can be compassionate is solved by the view—at this current stage of development—that algorithm programmers need to be compassionate. This does not mean that a machine cannot itself become compassionate in another sense. For instance, it can become compassionate if it exhibits the qualities of a compassionate being. Ultimately, it does not matter whether or not a machine is conscious in the normal sense. As long as the machine exhibits the outward characterization of interdependence and altruism, it can be said to be compassionate. I also argue that the ethics of AI must be integral to the coding of its program. In other words, the ethics—how we would like the AI to behave based on our own ethical beliefs—needs to be programmed into the AI software from the very beginning. I also reply to several objections against this idea. In essence, coding ethics into a machine does not imply that such ethics belongs solely to the programmer, nor does it mean that the machine is thereby completely estranged from its socio-cultural context.

### Policy Recommendations

1. **Programmers and software companies need to implement compassionate AI programs.** This is the key message from this article. No matter what kind of "social good" the AI is supposed to bring about, the software needs to be compassionate and ethical in the Buddhist sense.

2. **The public sector needs to ensure that rules and regulations are in place in order to create an environment that facilitates the development of ethical AI for social good.** Such rules and regulations will ensure that private companies have a clear set of directives to follow, and will create public trust in the works of the private sector.

# Moralizing and Regulating Artificial Intelligence: Does Technology Uncertainty and Social Risk Tolerance Matter in Shaping Ethical Guidelines and Regulatory Frameworks?

M. Jae Moon and Iljoo Park, Institute for Future Government, Yonsei University

Examining technology uncertainty and social risk in the context of disruptive technologies, this study reviews the development of ethical guidelines for AI developed by different actors as a loosely institutional effort to moralize AI technologies. Next, we specifically examine the different regulatory positions of four selected countries on autonomous vehicles (AVs). Based on the status of moralizing and regulating AI, several policy implications are presented as follows:

1. Moralizing disruptive technologies should precede, and should be fully discussed and shared among different stakeholder prior to regulating them. Before a society adopts and enacts specific regulatory frameworks for disruptive technologies, ethical guidelines (i.e., AI principles or AI ethical guidelines) must be jointly formulated based upon a thorough deliberation of particular disruptive technologies by different stakeholders representing industries, researchers, consumers, NGOs, international organizations, and policymakers.

2. AI ethical guidelines should support sustainable and human-centric societies by minimizing the negative socio-economic and international consequences of disruptive technologies (i.e., inequality, unemployment, psychological problems, etc.), while maximizing their potential benefits for environmental sustainability, quality of life among others.

3. Once a general consensus is made on general ethical guidelines, they should be elaborated and specified in details targeting individual stakeholder groups representing different actors and sectors.

Specific AI ethical guidelines should be developed and customized for AI designers, developers, adopters, users, etc. based on the AI lifecycle. In addition, industry and sector specific ethical guidelines should be developed and applied to each sector (care industry, manufacturing industry, service industry, etc.).

4. In regulating AI and other disruptive technologies, governments should align regulations with key values and goals embedded in various AI ethical guidelines (transparency, trustworthiness, lawfulness, fairness, security, accountability, robustness, etc.) and aim to minimize the potential social risks and negative consequences of AI by preventing and restricting possible data abuses or misuses, ensuring fair and transparent algorithms, in addition to establishing institutional and financial mechanisms through which the negative consequences of AI are systematically corrected.

5. Governments should ensure the quality of AI ecosystems by increasing government and non-government investment in R&D and human resources for AI by maintaining fair market competition among AI-related private companies, and by promoting AI utilities for social and economic benefits.

6. Governments should carefully design and introduce regulatory sandbox approaches to prevent unnecessarily strict and obstructive regulations that may impede AI industries but also facilitate developing AI and exploring AI-related innovative business models.

# Definition and Recognition of AI and its Influence on the Policy: Critical Review, Document Analysis and Learning from History

Kyoung Jun Lee, School of Management, Kyung Hee University
Yujeong Hwangbo, Department of Social Network Science, Kyung Hee University

## Abstract

Opacity of definitions hinders policy consensus; and while legal and policy measures require agreed definitions, to what artificial intelligence (AI) refers has not been made clear, especially in policy discussions. Incorrect or unscientific recognition of AI is still pervasive and misleads policymakers. Based on a critical review of AI definitions in research and business, this paper suggests a scientific definition of AI. AI is a discipline devoted to making entities (i.e., agents and principals) and infrastructures intelligent. That intelligence is the quality which enables entities and infrastructures to function (not think) appropriately (not humanlike) as an agent, principal, or infrastructure. We report that the Organisation for Economic Co-operation and Development (OECD) changed its definition of AI in 2017, and how it has since improved it from "humanlike" to "rational" and from "thinking" to "action". We perform document analysis of numerous AI-related policy materials, especially dealing with the job impacts of AI, and find that many documents which view AI as a system that "mimics humans" are likely to over-emphasize the job loss incurred by AI. Most job loss reports have either a "humanlike" definition, "human-comparable" definition, or "no definition". We do not find "job loss" reports that rationally define AI, except for Russell (2019). Furthermore, by learning from history, we show that automation technology such as photography, automobiles, ATMs, and Internet intermediation did not reduce human jobs. Instead, we confirm that automation technologies, as well as AI, creates numerous jobs and industries, on which our future AI policies should focus. Similar to how machine learning systems learn from valid data, AI policymakers should learn from history to gain a scientific understanding of AI and an exact understanding of the effects of automation technologies. Ultimately, good AI policy comes from a good understanding of AI.

## Policy Recommendations

1. Policy experts should be well educated about what AI is and what is really going on in AI research and business. Specifically, AI should be considered a discipline that allows entities and infrastructures to become intelligent. This intelligence is the quality that enables agents, principals, and infrastructures to function appropriately. AI should not be considered a humanlike or super-human system. As such, previous AI policies based on the old paradigm should be rewritten.
2. Governments should create programs to educate administrative officials, policy experts in public-owned research institutes, and lawmakers in national assemblies.
3. Similar to how machine learning systems learn from valid data, policymakers should learn from history, as well as recognize the positive impacts of automation technology. New AI policies should then be established based on this new recognition.
4. When adopting AI, governments and society should recognize its characteristics as an optimization system in order to create more public benefit, faster business outcomes, and less risk.

# Regulatory Interventions for Guiding and Governing the Use of Artificial Intelligence by Public Authorities

Arindrajit Basu, Elonnai Hickok and Amber Sinha, Centre for Internet & Society, India

## Summary

The use of artificial intelligence (AI)-driven decision-making in public functions has been touted around the world as a means of augmenting human capacities, removing bureaucratic fetters, and benefiting society. This certainly holds true for emerging economies. Due to a lack of government capacity to implement these projects in their entirety, many private sector organizations are involved in traditionally public functions, such as policing, education, and banking. AI-driven solutions are never "one-size-fits-all" and exist in symbiosis with the socio-economic context in which they are devised and implemented. As such, it is difficult to create a single overarching regulatory framework for the development and use of AI in any country, especially those with diverse socio-economic demographics like India. Configuring the appropriate regulatory framework for AI correctly is important. Heavy-handed regulation or regulatory uncertainty might act as a disincentive for innovation due to compliance fatigue or fear of liability. Similarly, regulatory laxity or forbearance might result in the dilution of safeguards, resulting in a violation of constitutional rights and human dignity. By identifying core constitutional values that should be protected, this paper develops guiding questions to devise a strategy that can adequately chart out a regulatory framework before an AI solution is deployed in a use case. This paper then goes on to test the regulatory framework against three Indian use cases studied in detail – predictive policing, credit rating, and agriculture.

## Key Recommendations

1. To adequately regulate AI in public functions, regulation cannot be entirely "responsive" as the negative fall out of the use case may be debilitating and greatly harm constitutional values. We therefore advocate for "smart regulation" – a notion of regulatory pluralism that fosters flexible and innovative regulatory frameworks by using multiple policy instruments, strategies, techniques, and opportunities to complement each other.
2. The five key values that must be protected by the state across emerging economies are: (1) agency; (2) equality, dignity, and non-discrimination; (3) safety, security and human impact; (4) accountability, oversight, and redress; and (5) privacy and data protection.
3. The scope, nature, and extent of regulatory interventions should be determined by a set of guiding questions, each of which has implications for one or more of constitutional values.
4. Whenever the private sector is involved in a "public function", either through a public–private partnership or in a consultation capacity, clear modes, frameworks, and channels of liability must be fixed through uniform contracts. The government may choose to absorb some of the liability from the private actor. However, if that is the case, this must be clearly specified in the contract and clear models of grievance redressal should be highlighted.
5. The case studies point to a need for constant empirical assessment of socio-economic and demographic conditions before implementing AI-based solutions.

6. Instead of replacing existing processes in their entirety, decision-making concerning AI should always look to identify a specific gap in an existing process and add AI to augment efficiency.
7. The government must be open to feedback and scrutiny from private sector and civil society organizations, as that will foster the requisite amount of transparency, trust, and awareness regarding the solution – all of which are challenges in emerging economies.
8. In situations where the likelihood or severity of harm cannot be reasonably ascertained, we recommend adopting the precautionary principle from environmental law and suggest that the solution not be implemented until scientific knowledge reaches a stage where it can reasonably be ascertained.

| VALUE | QUESTIONS |
|---|---|
| AGENCY | Is the adoption of the solution mandatory? |
| | Does the solution allow for end-user control? |
| | Is there a vast disparity between primary user and impacted party? |
| EQUALITY, DIGNITY, AND NON-DISCRIMINATION | Is the AI solution modelling or predicting human behavior? |
| | Is the AI solution likely to impact minority, protected, or at-risk groups? |
| SAFETY, SECURITY, AND HUMAN IMPACT | Is there a high likelihood or high severity of potential adverse human impact as a result of the AI solution? |
| | Can the likelihood or severity of adverse impact be reasonably ascertained with existing scientific knowledge? |
| ACCOUNTABILITY, OVERSIGHT, AND REDRESS | To what extent is the AI solution built with "human-in-the-loop" supervision prospects? |
| | Are there reliable means for retrospective adequation? |
| | Is the private sector partner involved with either the design of the AI solution, its deployment, or both? |
| PRIVACY AND DATA PROTECTION | Does the AI solution use personalized data, even in anonymized form? |

# AI Technologies, Information Capacity, and Sustainable South World Trading

Mark Findlay, Singapore Management University, School of Law – Centre for AI and Data Governance

## Abstract

This paper represents a unique research methodology for testing the assumption that AI-assisted information technologies can empower vulnerable economies in trading negotiations. Its social good outcome is enhanced through additionally enabling these economies to employ the technology for evaluating more sustainable domestic market protections. The paper is in two parts; the first presents the argument and its underpinning assumption that information asymmetries jeopardize vulnerable economies in trade negotiations and decisions about domestic sustainability. We seek to use AI-assisted information technologies to upend situations where power is the discriminator in trade negotiations because of structural information deficits, and where the outcome of such deficits is the economic disadvantage of vulnerable stakeholders. The second section is a summary of the empirical work piloting a more expansive engagement with trade negotiators and AI developers. The empirical project provides a roadmap for policymakers to adopt model reflections from focus groups and translate these into a real-world research experience. The research method has three phases, designed to include a diverse set of stakeholders – a scoping exercise, a solution exercise, and a strategic policy exercise. The empirical achievement of this paper is validating the proposed action-oriented methodology through a "shadowing" pilot device, where representative groups engaged their role-plays and represented essential understandings. General findings from the two focus groups are provided.

## Principal Policy Projections

- At the initiation of the project, **an intensive needs analysis should be initiated**, grounded in developing local skills around what questions to ask regarding information deficit, then translating into learning about what format to store and order data, and what data can accomplish in trading negotiations and domestic market sustainability. This exercise will empower domestic counterparts and achieve ownership. This exercise should be a collaboration between ESCAP, sponsor companies, and agencies;
- **Trading information asymmetries should be addressed** by sponsor companies, donors, and associated international agencies, through AI-assisted technologies for domestically empowering information access capacity building. **UN ESCAP should promote the use of AI-assisted technologies** to flatten information asymmetries that exist among trading partners in the region;
- While AI has the potential for empowering presently disadvantaged economies to negotiate in equal terms to raise the well-being of all people, such empowerment will not materialize without **adequate assistance**, in the form of technology, training, and domestic policy advice;
- **Product sustainability** is essential for the success of the project ongoing. Sponsor companies, and ESCAP in oversight, should ensure certain crucially sustainable deliverables covering: *data sources, data integrity and validation, accountability, and the technical sustainability of technical products.* These issues require allied services from sponsors, providers, advisers, and locally trained experts.

# Governing Data-driven Innovation for Sustainability: Opportunities and Challenges of Regulatory Sandboxes for Smart Cities

Masaru Yarime, Division of Public Policy, The Hong Kong University of Science and Technology

## Abstract

Data-driven innovation plays a crucial role in tackling sustainability issues. Governing data-driven innovation is a critical challenge in the context of accelerating technological progress and deepening interconnection and interdependence. AI-based innovation becomes robust by involving the stakeholders who will interact with the technology early in development, obtaining a deep understanding of their needs, expectations, values, and preferences, and testing ideas and prototypes with them throughout the entire process. The approach of regulatory sandboxes plays an essential role in governing data-driven innovation in smart cities, which faces a difficult challenge of collecting, sharing, and using various kinds of data for innovation while addressing societal concerns about privacy and security. How regulatory sandboxes are designed and implemented can be locally adjusted, based on the specificities of the economic and social conditions, to maximize the effect of learning through trial and error. Regulatory sandboxes need to be both flexible to accommodate the uncertainties of innovation, and precise enough to impose society's preferences on emerging innovation, functioning as a nexus of top-down strategic planning and bottom-up entrepreneurial initiatives. Data governance is critical to maximizing the potential of data-driven innovation while minimizing risks to individuals and communities. With data trusts, the organizations that collect and hold data permit an independent institution to make decisions about who has access to data under what conditions, how that data is used and shared and for what purposes, and who can benefit from it. A data linkage platform can facilitate close coordination between the various services provided and the data stored in a distributed manner, without maintaining an extensive central database. As the provision of personal data would require the consent of people, it needs to be clear and transparent to relevant stakeholders how decisions can be made in procedures concerning the use of personal data for public purposes. The process of building a consensus among residents needs to be well-integrated into the planning of smart cities, with the methodologies and procedures for consensus-building specified and institutionalized in an open and inclusive manner. As application programming interfaces (APIs) play a crucial role in facilitating interoperability and data flow in smart cities, open APIs will facilitate the efficient connection of various kinds of data and services.

## Policy Recommendations

1. Data governance of smart cities should be open, transparent, and inclusive to facilitate data sharing and integration for data-driven innovation while addressing societal concerns about security and privacy.
2. The procedures for obtaining consent on the collection and management of personal data should be clear and transparent to relevant stakeholders with specific conditions for the use of data for public purposes.
3. The process of building a consensus among residents should be well-integrated into the planning of smart cities, with the methodologies and procedures for consensus-building specified and institutionalized in an open and inclusive manner.
4. APIs should be open to facilitate interoperability and data flow for efficient connection of various kinds of data and sophisticated services in smart cities.

# Including Women in AI-enabled Smart Cities: Developing Gender-inclusive AI Policy and Practice in the Asia-Pacific Region

Caitlin Bentley, Katrina Ashton, Brenda Martin, Elizabeth Williams, Ellen O'Brien, Alex Zafiroglu, and Katherine Daniell, 3A Institute, Australian National University

Smart city initiatives are widespread across the Asia-Pacific region. AI is increasingly being used to augment and scale smart city applications in ways that can potentially support social good. We critically reviewed the literature on two key AI applications for social good: increasing safety and security in public spaces through the use of facial recognition technology, and improving mobility through AI-enabled transportation systems including smart traffic lights and public transportation route optimization. We find that there is an urgent need to consider how best to include women in the design, development, management, and regulation of AI-enabled smart cities. After all, poorly designed or delivered AI-enabled smart city technology could potentially negatively and differentially impact women's safety, security, and mobility. To address these pitfalls, we conducted interviews with a range of female and feminist scholars, activists, and practitioners – many of whom are working in the technology space. We carried out an analysis using the 3A Framework. This Framework focuses on investigating smart city initiatives through the themes of agency, autonomy, assurance, interfaces, indicators, and intent. We suggest the following actions be required: (1) commit to gender inclusive policymaking and praxis in national smart city policy; (2) institute formal consultation and participatory processes involving diverse women and community representatives through all stages of a smart city initiative; and (3) devise clearer roles and responsibilities surrounding the protection and empowerment of women in AI-enabled smart city initiatives.

1. **Commit to gender inclusive policymaking and praxis in national smart city policy:** High-level national smart city documentation frequently makes reference to social inclusion goals, but little is mentioned on how social inclusion is practiced. AI-enabled smart cities involve an interlaced network of actors, such as government ministries, private sector actors, and community groups.

Governments can play a key coordination role, whilst guiding the establishment of common goals and practices. Moreover, countries across Asia-Pacific should review national policy to take into account the interconnected nature of smart city initiatives, and how they connect to multiple targets across the Sustainable Development Goals (SDGs). National governments should institute a process to develop indicators that map smart city progress in the pursuit of achieving SDGs, namely SDG 5 and 11.

2. **Institute formal consultation and participatory processes involving diverse women and community representatives through all stages of a smart city initiative:** Our research identifies new models of design, community ownership, and public debate supported by AI. Municipal actors, industry partners, and women's community groups should invest greater resources into experimenting with innovative engagement and representation models, as well as building into project plans the time needed for engagement. The 3A Framework can be used to guide discussions with communities, women, and their representatives. Our research highlights how the Framework sheds lights on multiple and interrelated systemic factors that need to be taken into consideration, rather than focusing only on the perspectives of individuals.

3. **Devise clearer roles and responsibilities surrounding the protection and empowerment of women in AI-enabled smart city initiatives:** There is an urgent need for policymakers to establish greater transparency and clearer rules around the handling, ownership, and protection of data with, for, and about women. Better understanding of the impacts, not only the performance of these systems, should guide this discussion. Consequences for mistreatment, harm, and mismanagement across all levels of smart city initiatives should be carefully and clearly outlined. More opportunities for women to be consulted and involved in the design, management, evaluation, and regulation of AI-enabled smart city initiatives are warranted.

# AI and the Future of Work: A Policy Framework for Transforming Job Disruption into Social Good for All

Wilson Wong, Chinese University of Hong Kongy

## Abstract

This paper examines the impact of artificial intelligence (AI) on the future of work to develop a policy framework for transforming job disruption caused by AI into social good for all. While there is a considerable amount of research and discussion on the impact of AI on employment, there is relatively less research on what governments should do to turn the risk and threat of AI into job opportunities and social good for all. This paper consists of two major parts. It first builds on the typology of job replacement and AI to establish a policy framework on the role of the government, as well as the policy responses it should make to address various concerns and challenges. On the principle of "rise with AI, not race with it", the government must play an active or even aggressive role not only for retraining knowledge, skill-building, and job re-creation, but also for social security and a fair re-allocation of resources in the job disruption process. Second, the paper conducts a survey of national AI strategies to assess the extent to which AI policy of job disruption is addressed by other countries. It concludes that many countries, especially developing ones, are not well-prepared for AI, and most countries seem to be overlooking fairness and equity issues under job disruption in the arrival of the AI era.

## Policy Summary: Major Recommendations

1. Theory and Practice: Governments should have more alignment and integration between theory and policy in formatting their AI strategies. For example, they should discuss how enabling technologies as well as social and creative intelligence are included in their retraining, reskilling, and education programs.
2. International Organization and Developing World: AI impacts on both developed and developing worlds. Many developing countries are ill-prepared due to limitations in resources and other factors. International organizations such as the United Nations (UN) should offer more support to these nations to help set up their own AI strategies to evaluate threats and opportunities and formulate solutions.
3. AI for All (No One Left Behind): Equity, social security, and fair re-distribution, such as introducing Universal Basic Income (UBI) to protect vulnerable populations, are the missing pieces in the AI strategies of most countries. Governments should confront these important issues head on and incorporate them explicitly in their national AI strategies.

# Appendix 2

# Project History

The AI for Social Good Project is the heir to two series of policy advocacy initiatives on the digital economy by the Association of Pacific Rim Universities (APRU). The first series is the Digital Economy initiative and its successor, the AI for Everyone project, hosted by Keio University. The second series, led by The Hong Kong University of Science and Technology, is "Transformation of Work in Asia Pacific in the 21st Century: Key Policy Implications".

The project also stems from the partnership UN ESCAP has been building with ARTNET on STI Policy – a regional research and training network supporting policy research to leverage science, technology, and innovation as powerful engines for sustainable development in Asia Pacific.

In addition to the authors represented in this project, the following advisory board members, to whom we are extremely grateful for their valuable input, were chosen to provide feedback about the projects.

| Name | Affiliation |
|---|---|
| Hideaki Shiroyama | The University of Tokyo |
| Pascale Fung | The Hong Kong University of Science and Technology |
| Toni Erskine | Australia National University |
| Yudho Giri Sucahyo | University of Indonesia |
| P. Anandan | Wadhwani Institute of AI, Mumbai |
| Hoyoung Lee | Korea Information Society Development Institute |
| Punit Shukla | World Economic Forum |
| Yongyuth Yuthavong | National Science and Technology Development Agency |

Table 1: List of advisory board members

To kick-off this collaborative project, the first face-to-face meeting was held on June 5, 2019 at Keio University's Mita campus. A virtual policy fora for the dissemination and discussion of project findings is planned to be held later in the year.

One last face-to-face meeting before final submission of the output, together with an open-to-public forum, was originally scheduled for February 20 – 21, 2020. However, due to the COVID-19 pandemic, it was replaced by an online meeting of just the project members. The project outputs were submitted in May 2020 for editing and subsequent publication in August 2020. When it is safe to do so, an open-to-public forum will be held.
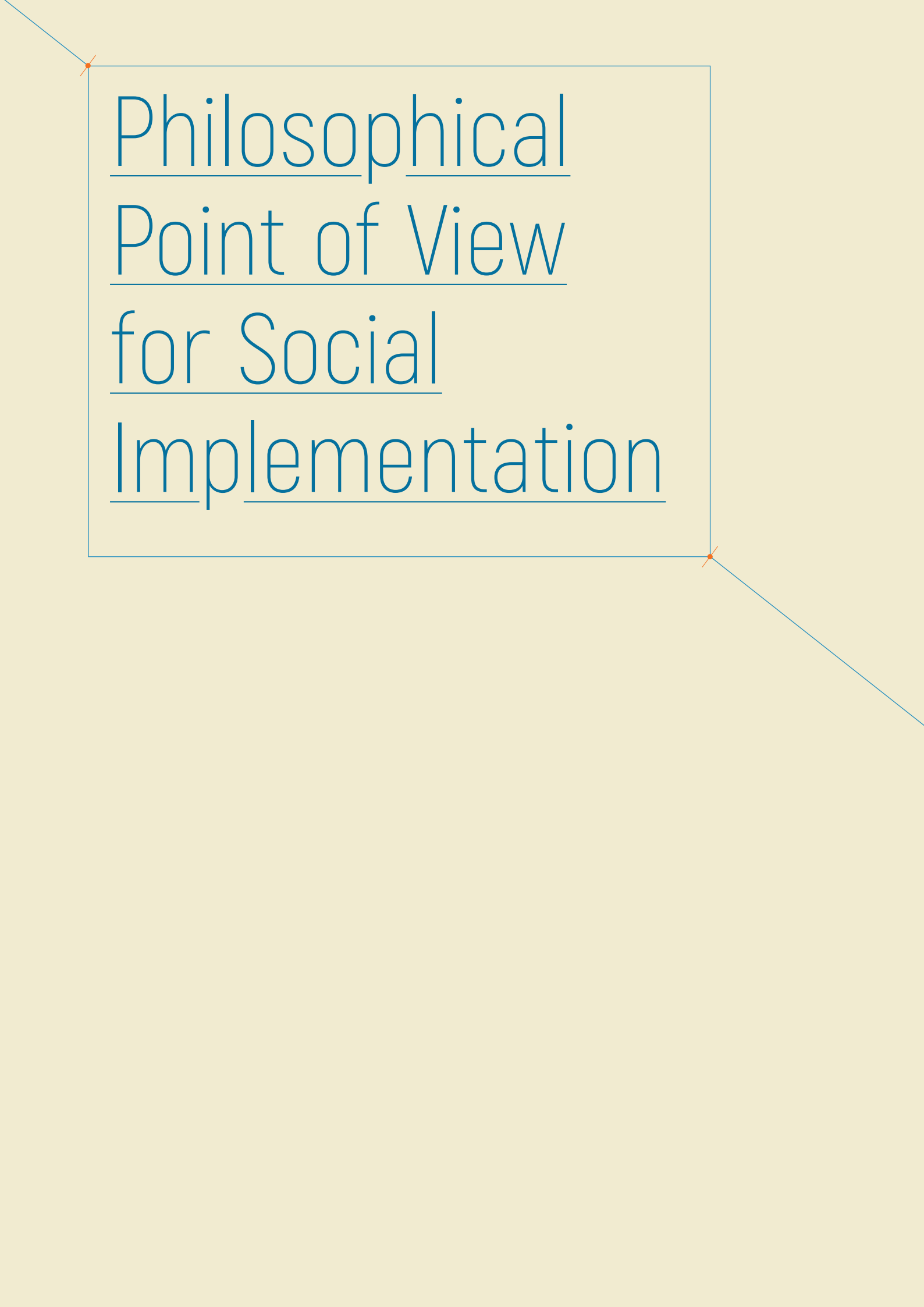
The project was organized by the following members:

| Name | Affiliation |
|---|---|
| Jiro Kokuryo, Project Coordinator | Keio University |
| Yoshiaki Fukumi | Keio University |
| Cherry Wong | Keio University |
| Daum Kim | Keio University |
| Minkyoung Cho | Keio University |
| Christina Schönleber | APRU |
| Tina Lin | APRU |
| Sanghyun Lee | Google |
| Jake Lucchi | Google |
| Marta Perez Cuso | UN ESCAP |

**Table 2: Organizing members**

We are grateful for all the efforts of those involved and sincerely hope that this document will help policymakers in the region accomplish their goals.

# Philosophical Point of View for Social Implementation

# AI for
# Social Good:

# Buddhist
# Compassion
# as a Solution

**Soraj Hongladarom**
Department of Philosophy,
Faculty of Arts,
Chulalongkorn University

## Abstract

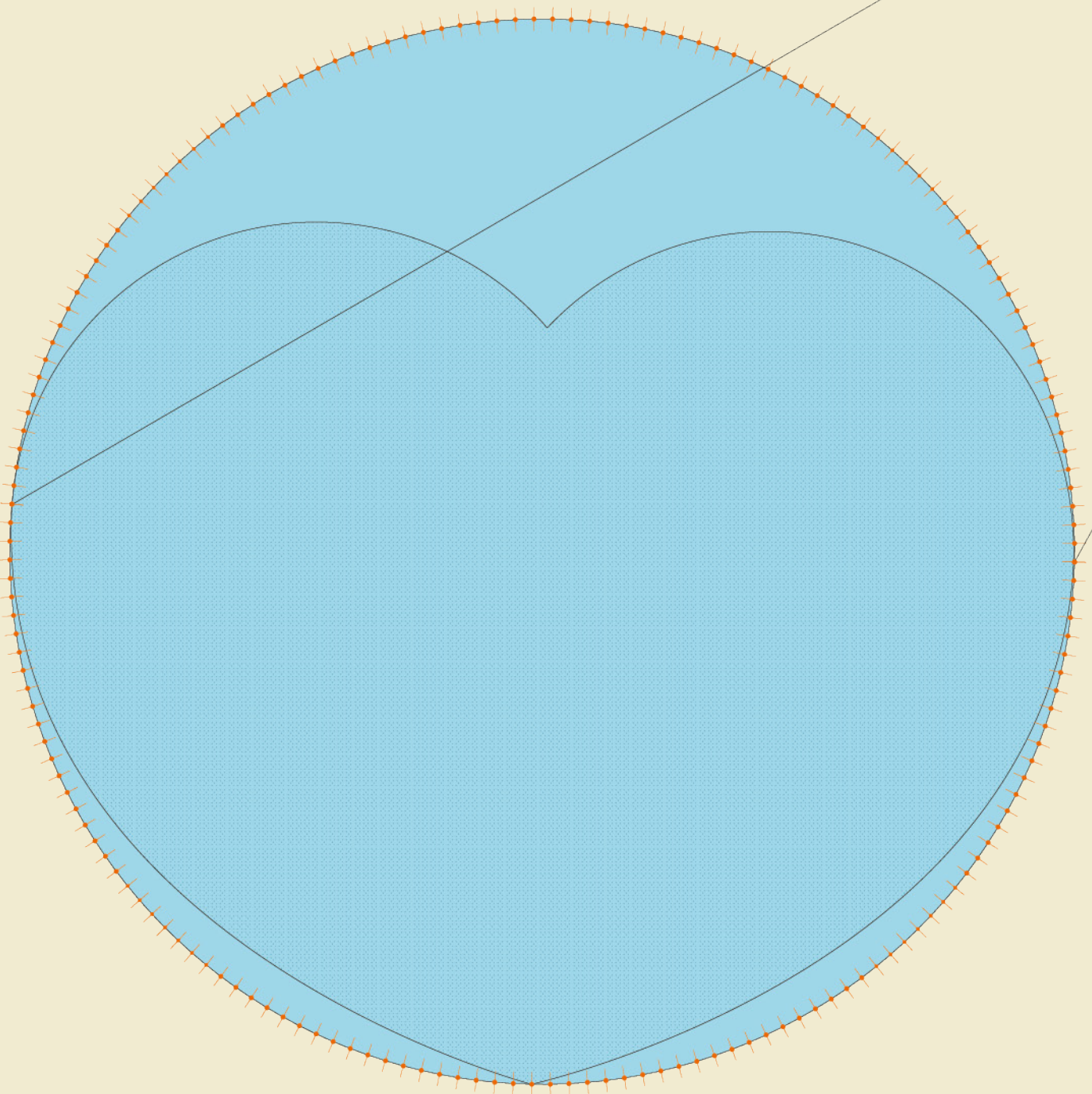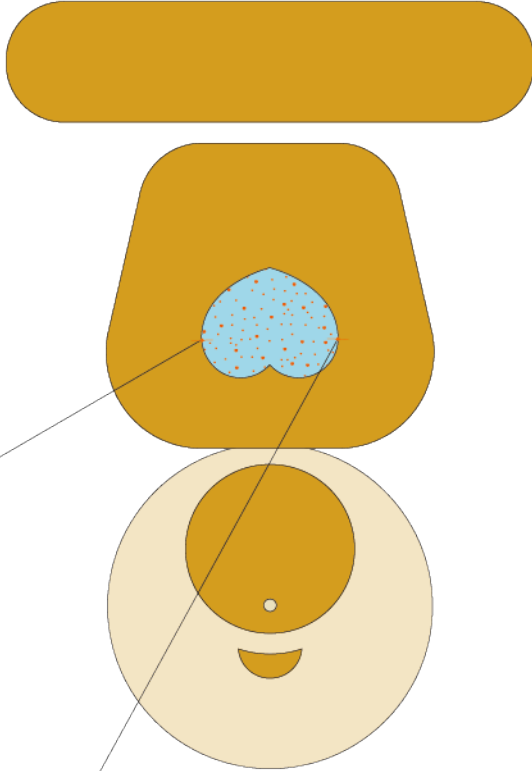In this paper, I argue that in order for artificial intelligence (AI) to deliver social good, it must be ethical first. I employ the Buddhist notion of compassion (*karunā*) and argue that for anything to be ethical, it must exhibit qualities that characterize compassion, namely the realization that everything is interdependent and the commitment to alleviate suffering in others. The seemingly incoherent notion that a thing (e.g., an AI machine or algorithm) can be compassionate is solved by the view that – at this current stage of development – algorithm programmers need to be compassionate. This does not mean that a machine cannot become compassionate in another sense. For instance, a machine can become compassionate if it exhibits the qualities of a compassionate being, regardless of whether it is conscious. As long as the machine exhibits the outward characterization of interdependence and altruism, then it can be said to be compassionate. This paper also argues that the ethics of AI has to be integral to the coding of its program. In other words, the ethics (i.e., how we would like the AI to behave based on our ethical standpoint) needs to be programmed into the AI software from the very beginning. This study has also replied to several objections against this idea. To summarize, coding ethics into a machine does not imply that the ethics thus coded belongs solely to the programmer, nor does it mean that the machine is thereby completely estranged from its socio-cultural context.

## Introduction

In the past few years, few innovations in technology have aroused as much public interest and discussion as AI. After many years of lying in the doldrums, with many broken promises in the past decades, AI once again became a focal point after it defeated both the European champion and reigning world champion at the ancient game of Go in 2016. The defeat was totally unexpected, as computer scientists and the public believed that Go was much more complex than chess. Since the number of possible moves that needed to be calculated was too vast for any computer to calculate, many believed that Go represented the supreme achievement of human beings, and could not be bested or emulated by a machine. Thus,

there was worldwide sensation after both the European champion Fan Hui, and Lee Sedol, the world champion, were soundly defeated at Go by a machine in a relatively short span of time. Following this AI victory, it became clear that no human could ever defeat a machine in a board game.

What ensued was an explosion in the power of AI — a resurgence after many years of dormancy and repeated failed promises. AI has been with us for many decades. Computer scientists who developed it believed that a computer could actually mimic the workings of the human brain. The project seemed promising at first; for example, the computers could play Tic-Tac-Toe, Checkers, and eventually chess. Some progress was also made in the field of natural language processing and machine translation. Nonetheless, these successes were not as spectacular as the scientists themselves had envisioned, and AI was unable to fulfil the expectations that its developers had originally claimed. For example, the expert system environment was developed during the early 1980s, but was prone to mistakes and thus became not suitable for normal use. The market for expert systems thus largely failed. Many promises of AI systems at that time, such as speech recognition, machine translation, and others, were not fulfilled. As a result, funding was largely cut, and AI research made very little progress. These failures were largely due to the fact that computers at that time lacked power, and data, so their predictive power remained limited.

The software that created history, AlphaGo, was developed by DeepMind, a British company founded in 2010 and acquired by Google in 2014. The company made history in 2015 and 2016 when its AI creation, AlphaGo, defeated both the European champion and the world champion of Go. The technique used by AlphaGo was radically different from Deep Blue, a software developed by IBM which defeated the chess world champion, Gary Kasparov, in 1997. Deep Blue used GOFAI, or "good old-fashioned AI", to blindly search for the best possible moves using a brute force search technique. This technique proved unfeasible for much more complex games such as Go, where the number of possible moves exceed the number

of atoms in the universe. Thus, AlphaGo used a new technique which was also being developed at that time. The new technique, known as deep learning, avoided the brute force search technique, and instead relied on very large amounts of data. The program learned from this data to determine the best moves. The data from millions of past moves made by humans limited the number of possible moves that the algorithm would need to make, thus enabling it to focus on the most relevant moves. This, coupled with more powerful hardware, contributed to the program defeating Lee Sedol. The event was watched by many people worldwide, and its success was a "Sputnik moment" in terms of bringing AI back into the spotlight. Now, many researchers are racing against each other to find the most useful applications for the technology.

Many applications are being touted as potential ways in which deep learning AI could help to solve the world's problems. The following applications are currently being promoted: self-driving cars, deep learning (AI use) in healthcare, voice search or voice assistants, adding sounds to silent movies, machine translation, text generation, handwriting generation, image recognition, image caption generation, automatic colorization, advertising, earthquake prediction, brain cancer detection, neural networks in finance, and energy market price forecasting (Mittal, 2017). Some of these applications indeed address serious matters, such as self-driving cars and image recognition, while others are rather quaint, such as colorization or automatic sound generation in silent movies. In any case, Mittal mentions that some of the most prominent applications of deep learning (or machine learning) AI has emerged over the past three or four years. One of the most powerful uses of today's AI is its predictive power. Using vast data sources, AI promises to make predictions that would not be conceivable by human analysts. One of the promises, for example, concerns an AI system that can detect the onset of cancer by analyzing images of those who are still healthy. In other words, the power of today's AI lies in its ability to "see" things that are often undetected by trained specialists. The algorithm gains this ability through its analysis of

extensive data points that are fed into its system. The machine analyzes these data and finds patterns and correlations to make predictions.

This new technology has led many to look for ways in which AI could improve society. The applications mentioned in Mittal's article identifies some of the potential uses, or "social goods" that could be delivered by AI. Many large corporations have also jumped on the bandwagon in search of AI opportunities. Google, for example, has founded an initiative titled "AI for Social Good" (http://ai.google/social-good/), which aims at "applying AI to some of the world's biggest challenges", such as forecasting floods, predicting cardiac events, mapping global fishing activity, and so on (AI for Social Good, 2020).

This paper analyzes some of the ethical concerns arising from such applications. Researching the potential of AI to solve these problems is important, but when the technology is applied in real-world scenarios, care must be taken to ensure that the social and cultural environment is fully receptive to the technology. Not being receptive to the imported technology can lead to a sense of alienation, which can happen when the local population is excluded from the process of decision making regarding the adoption of the technology in question (Hongladarom, 2004). This could also lead to a resistance to AI technology. For example, using AI to forecast floods may lead to administrative measures that could cause mistrust or misunderstandings if the AI technology is not made clear to those affected by the measure. It is one thing for AI (if reliable) to identify when and where a severe flood will take place; it is another to convince a local population that a flood will occur and that their location will be affected. This shows that any successful employment of AI must factor in local beliefs and cultures. Moreover, the forecasting must not be used to gain an unfair advantage over others. For example, forecast knowledge of floods in a particular area and time might lead to hoarding or other unfair measures designed to maximize the individual gains of certain parties. This shows that ethics must always be integral to any kind of deployment of technology and its products.

Consequently, this paper aims to find ways in which machine learning AI could deliver social good in an ethical manner. More specifically, this paper argues that in order for AI to deliver social good, it must be ethical first. Otherwise, it might lead to negative outcomes that are similar to the aforementioned scenario of flood forecasting and hoarding. This is a vital principle to address, as sophisticated technology, such as facial recognition software, could be used to endanger people's right to privacy. As mentioned above, AI algorithms that forecast flooding could be used to gain unfair advantages over others. Hence, there must be a way for these algorithms themselves to act as safeguards against such use. For flood forecasting software, this might not be immediately apparent as it does not typically involve autonomous action. The software would likely deliver information and forecasting, with humans ultimately being responsible for acting on the information. However, even in this case, the software itself must be ethical on its own. At the very least, there should be some form of mechanism in which the possibility of misuse or abuse by certain groups (such as those intent on using the information to hoard food and other supplies) is minimized; such a mechanism should be installed as part of the software from the very beginning. Regarding facial recognition technology, the same type of mechanism should also be installed to avoid potential misuse. Simply, AI should be an integral part of an ethical way of living, right from the moment of implementation. Hence, instead of regarding AI and its surrounding technologies as something imported and inherently harmful towards the developing world, we must find a way in which AI becomes integral to help these people flourish.

Furthermore, this paper argues that the details of how to live an ethical life should include insights obtained from Buddhism; specifically, the teachings on compassion (*karunā*), which is one of the most important tenets of Buddhism. It may be suggested that Buddhist compassion — a concept that will be further developed in this paper — should play a key role in developing an *ethical* AI. This development then comprises the possibility of AI to deliver social good and function as an integral part of ethical living.

AI is undoubtedly powerful and has the potential to significantly change the world. Power always has to be accompanied by corresponding responsibility, restraint, and other ethical virtues.

The next section of this paper will review some of the current literature on the ethics of AI and AI for social good. Section 3 deals with the basic concepts of Buddhism. Section 4 presents the paper's main argument, together with replies to some of the objections during the course of research. The last section concludes with two main policy recommendations for the public sector and tech companies.

## AI for Social Good

The advent of AI has given rise to a plethora of ethical guidelines that aim to regulate AI research and development worldwide. A survey of the literature on AI for social good revealed that much of the literature overlaps with the ethics of AI and proposals for AI ethics guidelines in general. This is not surprising, as proposing AI for social good implies that AI should act ethically; by promoting social good, AI thereby becomes ethical. However, this transition is not automatic; one still has to provide an account of why it is indeed the case. The need for such an account seems to be more acute when an AI program might be created with the aim of providing a social good, but instead, turns out to be harmful. This justification forms one of the main objectives of this paper.

Nevertheless, it is important to review the literature on ethics guidelines for AI, as well as AI for social good, to provide a general outline and identify some of the key issues. A website titled "AI Ethics Guidelines Global Inventory" (**https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/**) has documented 82 guidelines. However, only four Asian countries are represented on the list: China, Korea, Dubai, and Japan. It should also be noted that none of the documents published in these countries are based on their own indigenous intellectual resources (see also Gal, 2019). This shows that there is a very high level of interest in how AI should be ethically grounded. In a related paper, "The Ethics of AI Ethics", Thilo Hagendorff

(Hagendorff, 2019) documents the ethical concepts that are mentioned in some of these guidelines, and identifies the top five concepts, which include privacy, accountability, fairness, transparency, and safety (Hagendorff, 2019). These factors largely correspond with a list in another paper written by Luciano Floridi and others (Floridi et al, 2020), where seven "essential factors" are listed, namely: (1) falsifiability and incremental deployment, (2) safeguards against manipulation of predictors, (3) receiver-contextualized intervention, (4) receiver-contextualized explanation and transparent purposes, (5) privacy protection and data subject consent, (6) situational fairness, and (7) human-friendly semanticization (Floridi et al, 2020, p. 5). Here, falsifiability means that the software system needs to be empirically testable, and only if it is testable will it be deemed trustworthy. Factor (2) (safeguards against predictors) is rather straightforward; it means that there needs to be a mechanism whereby false manipulation of input into the software is prevented, so that the results produced by the software are not biased. Factor (3) (receiver-contextualized intervention) refers to respecting the autonomy of the user; any intervention performed by the software needs to be "contextualized" to the needs and desires of the user. Factor (4) (receiver-contextualized explanation and transparent purposes) refers to respecting the autonomy of the user in terms of the software being easy and transparent to understand, where nothing important is hidden. Factor (5) (privacy protection and data subject consent) is self-explanatory and is the number one concern in the guidelines studied in Hagendorff's paper. Factor (6) (situational fairness) refers to the need for the software to maintain objectivity and neutrality by avoiding data input that is biased from the beginning. Factor (7) (human-friendly semanticization) means that humans should still maintain a level of control when the software is allowed to interpret and manipulate meaningful messages. For example, AI software can create clearer communication between the caregiver and patient, without intervening and excluding the caregiver from the process (Floridi et al, 2020, pp. 5-19).

These factors and concepts are also very much related to another set of concepts, also developed primarily by Floridi (Floridi et al, 2018; see also Cowls and Floridi,

2018). In this paper, Floridi and his team delineate five elements that are necessary for "good" AI in society. Most of these elements resemble the familiar ethical principles found in other areas of applied ethics, most notably in medical ethics. These are beneficence, non-maleficence, autonomy, and justice. Then Floridi and his team add another factor, explicability, which is unique to AI as it tends to operate in a "black box", where the normal user has no clue over how it works and how it comes up with its own answers (Floridi et al, 2018). Moreover, Mariarosario Taddeo and Floridi also have another article published in Science in 2018 mentioning the need for these factors for a good AI society (Taddeo and Floridi, 2018). They also discuss the need for what they call a "translational ethics" that combines foresight methodologies and analyzes of ethical risks (Taddeo and Floridi, 2018). In addition, these five principles are also discussed in The European Commission's High Level Expert Group on Artificial Intelligence (The European Commission's High-Level Expert Group on Artificial Intelligence, 2018, pp. 8-10), with the emphasis that AI systems need to be "human-centric" (The European Commission's High-Level Expert Group on Artificial Intelligence, 2018, p. 14). The overall concern of the document is that AI needs to be "trustworthy", and the requirements discussed here are among the necessary conditions. More specifically, the document discusses ten factors that are supposed to be sufficient for a trustworthy AI system. These are accountability, data governance, design for all, governance of AI autonomy (human oversight), non-discrimination, respect for (and enhancement of) human autonomy, respect for privacy, robustness, safety, and transparency (The European Commission's High-Level Expert Group on Artificial Intelligence, p. 14). Thus, these ten requirements largely mirror the requirements or essential factors mentioned earlier. Chief among these lists are factors such as autonomy, privacy, safety, and transparency. It is clear that there are many overlaps among such guidelines, with only relatively small differences among them.

Furthermore, Ben Green (Green, 2019) argues that computer scientists cannot rely on the idea that algorithms alone can solve the world's problems, but they need to see how social programs (which AI for Social Good is supposed to solve) are all connected with deeper and more intricate interconnections, which mere technical means alone cannot solve. Bettina Berendt, in a similar vein, proposes an "ethics pen-testing" where the design of AI is critically challenged by a series of questions aimed at the designer to defend himself/herself and to show that the design is ethically sensitive, all in order to improve the software design (Berendt, 2019). What is interesting in both Green's and Berendt's papers is that they are not content on merely proposing a list of guidelines for AI developers to follow, and instead point out that AI researchers and developers must be aware of ethics during all stages of development. Technical solutions alone are not enough, and will not be effective in bringing about the proposed "social good" of AI.

What has emerged is that most of the literature focuses on a list of ethical principles which, they argue, should be necessary for an effective ethical AI system. However, only a few works (e.g., Green and Berendt) argue that simply providing such a list bypasses the deeper interweaving connection between ethical principles and the underlying social and cultural contexts. Nonetheless, both Green and Berendt address these contexts in a vertical manner. More specifically, they focus on the interrelations between ethical principles and the wider concerns in a Western context. As mentioned earlier, there are only a few guidelines in Asia, and more interestingly, these guidelines do not mention their own intellectual resources. Hence, a large gap exists in the literature, namely the formulation of AI ethics principles based on the intellectual resources of the East. In fact, my recent book, "The Ethics of AI and Robotics: A Buddhist Viewpoint", discusses this issue in great detail (Hongladarom, 2020). Moreover, going beyond the gap in theoretical terms, there is also a gap in the content of the proposed guidelines. What I propose in this paper is that a complimentary principle of Buddhist ethics should be adopted as the foundation for thinking and deliberating on the ethics of AI and AI for social good. Furthermore, the principle of *karunā* (compassion) should be considered for the ethical guidelines of AI and for any theory related to AI for social good.

# Buddhist Ethics and Basic Buddhist Principles

It is not possible to explain all of the principles of Buddhism in this paper. Nonetheless, a very brief introduction to its relevant principles should provide a better context for the argument. More details on the principles of Buddhist ethics and an introduction to Buddhist philosophy can be found in the book that I mentioned earlier (Hongladarom, 2020). The book explains that Buddhist ethics is based on the idea that an action is considered right if it brings out something that is universally desired by all human beings, and wrong if it goes in the opposite direction. Thus, Buddhist ethics is markedly different from modern ethical theories; for instance, other theories do not specify what is universally desirable for all humans. In Immanuel Kant's ethical theory, for example, the basic idea of what constitutes a good action comes without considering the possible consequences of that action. Instead Kant's theory questions whether the action follows a universalizable maxim or not. The universally desirable goal, on the contrary, is definitely a goal; thus, Buddhist theory is in opposition with Kant's deontological theory. Furthermore, Buddhist theory is also different from utilitarianism in that, although utilitarianism is a kind of consequentialism, Buddhist theory specifies a definite content of the goal that is universally desirable to all human beings. Conversely, utilitarianism does not specify any definite content, and instead focuses on content that is deemed utilitarian. Buddhism suggests the possibility of a universally desirable goal that is valid for everyone. Since everyone desires happiness and wishes to avoid suffering, it may be seen as a universal goal. Buddhism has a very detailed theory regarding the definition of happiness as a universal goal. In simple terms, it describes a type of happiness that results when one's action is in total accordance with nature. Thus, the kind of "happiness" that results from indulging in sensual pleasure would not qualify, as this pleasure also brings about suffering. For example, eating certainly brings pleasure, but too much eating can cause a certain degree of discomfort, such as feeling bloated, etc. Therefore, true happiness (i.e., without suffering) is only attainable through a true understanding of nature.

This does not mean becoming a scientist, but instead, understanding that nature works according to the rules of cause and effect. Realizing this is a necessary step towards attaining what Buddhists call "*nirvāna*" or total cessation of all suffering. The term is usually translated as "Enlightenment." Hence, Buddhist ethical theory explains that an action is good if it leads to *nirvāna*, and vice versa.

As mentioned earlier, the aim of this paper is to show that Buddhist philosophy can contribute to the ethics of AI and AI for Social Good. A key point is that a person's actions must be in tune with nature. When this is the case, they essentially become one with nature. This is a concrete expression of the realization that there is no attachment to the ego, since it is just such an attachment that separates one from becoming fully in tune with nature. Compassion is a key ingredient in this realization, and what is truly good is the realization that there is no boundary between the ego and everything else, as well as the resultant desire to help others get rid of their suffering, which is ultimately due to a lack of realization. In the area of AI ethics and AI for social good, this means that one has to find a way in which AI can contribute to relieving the suffering of all beings. This may not be as grandiose as it may sound, as we are more than capable of finding out specific and concrete ways to achieve this. Doing so is to implement an ethics of AI that is in accordance with Buddhist ethical principles. The main idea being that, in order for AI to provide social good, it must consider the contexts involved, which may vary from place to place. A solution that might work in one context might not work in another. The examples put forward in this paper are flood forecasting and facial recognition, however, we can certainly imagine other cases. In the field of automated reasoning or decision making, one also needs to be careful that the decisions made by AI are always accountable to humans. Allowing AI to have a free hand in making decisions (such as in stock trading) would go against the Buddhist principle of compassion, as this tends to create more suffering rather than reduce it.
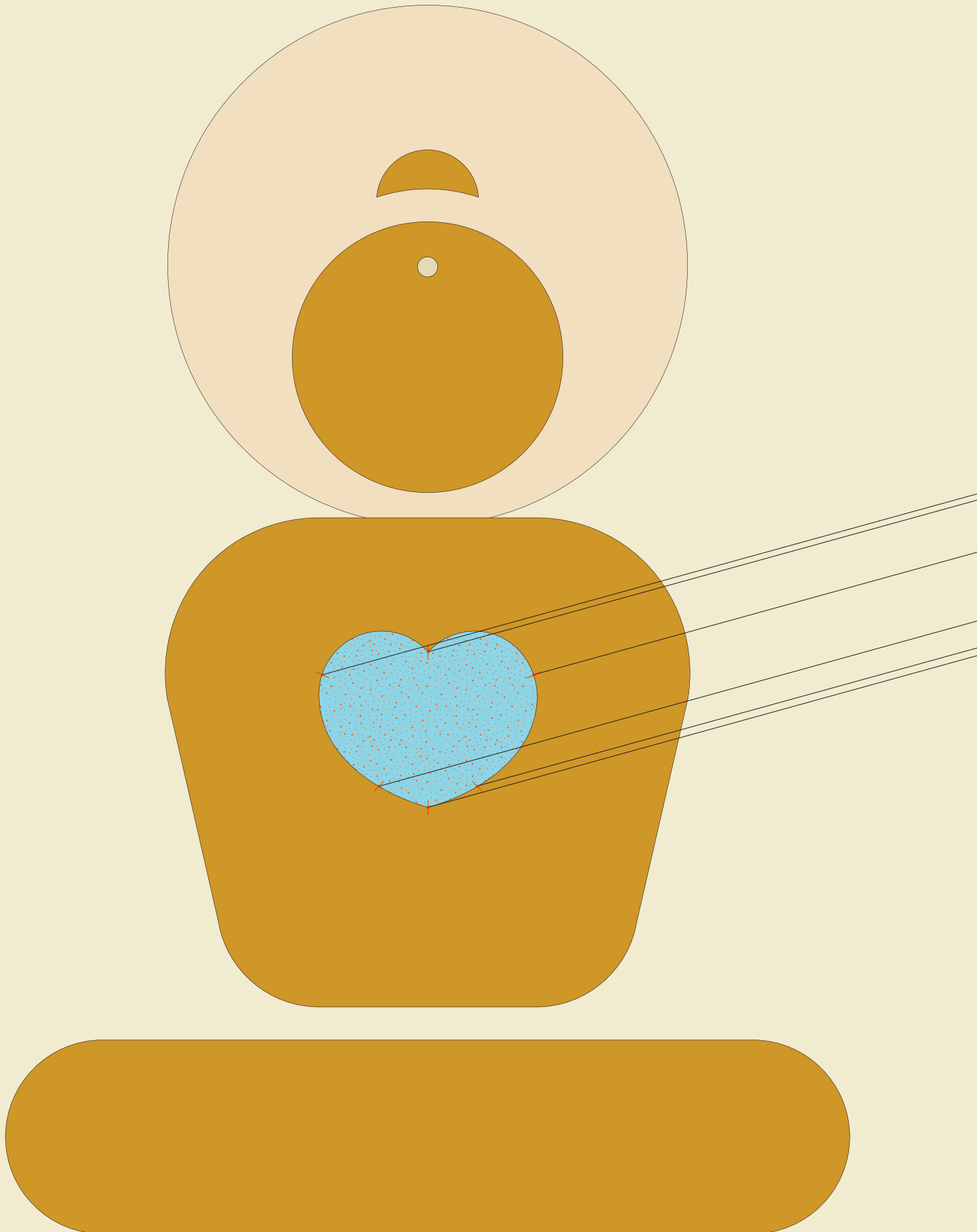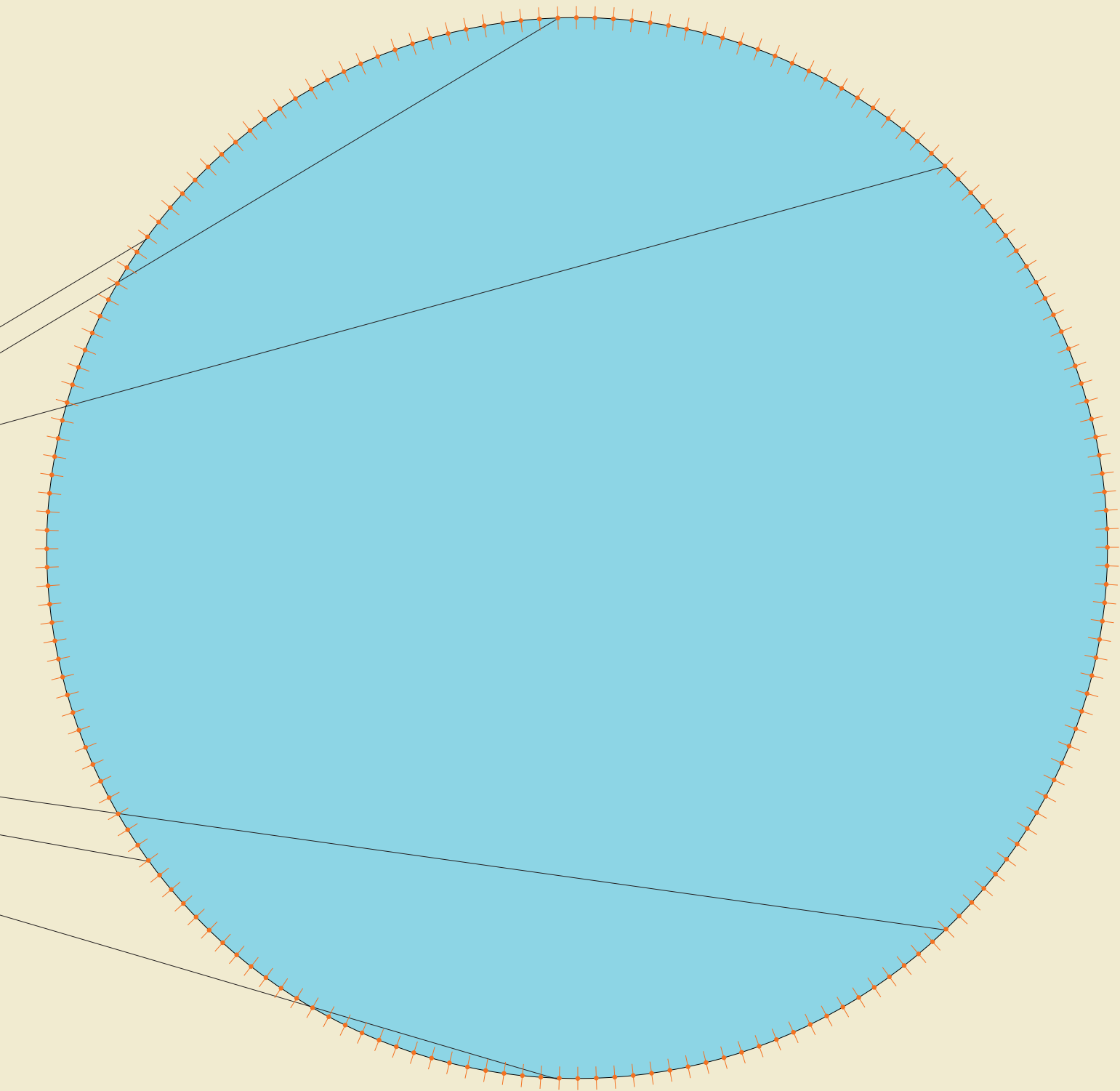
## Could AI Become Compassionate?

As we have seen, this study argues that AI needs to be compassionate. This means that AI must exhibit the two qualities that constitute compassion, namely interdependence and altruism. AI exhibits interdependence by showing concretely that it understands (within the constraints of current AI technology) the concept of things being interdependent and interconnected. This can be achieved with an AI algorithm that shows concern for the welfare of someone or something. For example, the aforementioned flood forecasting algorithm could show a level of understanding of interdependence by having in its internal mechanism connected to other relevant factors that are no less important, such as economic conditions, price forecasting, political climate, and geographical information, etc. AI flood forecasting could lead to the hoarding of essential food and supplies, which is an unethical act. However, the algorithm might struggle to learn how its predictions could be used by humans in a negative way. Here, a program that embeds algorithms in a larger context could make it more difficult for information to be used for personal gains. For example, the algorithm could publicly broadcast its predictions, making it impossible for certain parties to gain an advantage. An internal "safety lock" within the algorithm could be installed as an indelible component to make it imperative to broadcast information to everyone involved rather than to individual users. The broadcasting feature may, however, be necessary for flood forecasting, but broadcasting on this scale might be unethical in other contexts or for certain algorithms. For example, some algorithms are intended to work privately (e.g., personal health information). As such, developers need to see which contexts are relevant for installing safety mechanisms inside algorithms.

The other component of Buddhist compassion is the commitment to alleviate suffering for all sentient beings. Here, sentient beings are relieved of their suffering through someone who is completely compassionate. However, such an ideal is impossible to realize in reality, where the one who practices compassion has limited power. Nonetheless, we must do whatever we can—within the limits of our power—to help relieve suffering. For AI algorithms, this would mean taking active steps in creating a world where suffering is eliminated as much as possible. More specifically, the algorithm should be designed to help alleviate suffering from the very beginning. For example, facial recognition technology could be developed to recognize particular features so that certain traits are predicted, such as the onset of a disease, leading to early prevention. One may assume that suffering is unrelated to software development, as it appears to be an external requirement. However, it should be an integral part of software development in itself. This pertains to key areas or problems which AI algorithms will be designed to solve from the beginning.

Michael Kearns and Aaron Roth (Kearns and Roth, 2019) argue that an algorithm should be ethical in the sense that ethical components should be programmed into the algorithm. Here I suggest that compassion should also be programmed into AI algorithms. In fact, the same idea has already been proposed by James Hughes (Hughes, 2012). However, according to Hughes, a robot only becomes compassionate when it can imitate human emotion. I propose that compassion can be attained when it exemplifies the two components mentioned earlier, namely realization of interdependence and the commitment to relieve suffering. More specifically, a robot becomes compassionate when it exhibits genuine commitment and action geared toward alleviating suffering. Thus, it is more action-oriented than merely displaying or mimicking emotions.

How can we program robots or AI algorithms to be compassionate? We could say that an algorithm "understands" interdependence when it is programmed in such a way that it "recognizes" various external factors that are involved in making a more ethically nuanced assessment. Of course, the algorithm does not understand anything—we are not talking about a superintelligence—but it is a way of talking to show that the algorithm exhibits certain behaviors that we recognize colloquially as an

understanding. Hence, for the algorithm to understand interdependence, which is one component of Buddhist compassion, it has to exhibit certain external features that are not directly part of its core objective, so to speak. These features may not be part of the core mission, but they are very important in making an ethical judgment of the situation in which it is employed in order that it becomes more ethical. If a given objective, such as to maximize a certain output, is found to involve trade-offs between the output and other desirable factors, then the machine would be programmed not to follow the maximization. It will realize or "understand" that such an action leads to a contradiction with its own prime directive, which is to alleviate human suffering. To come back to flood prevention software, an algorithm might be taught to accurately predict floods in a certain area. However, predicting floods alone is not ethical as it could lead to hoarding, as we have seen. Thus, the AI needs to be programmed with compassion so that it can predict floods while also considering other relevant factors. For example, the AI could display a warning sign if a user attempts to misuse the data. Then, the second component of compassion, altruism, is ideally put into action when the algorithm initiates an action designed to help relieve affected persons from suffering. To use another example, a microloan algorithm might override its directive (maximizing profit for its creator or owner) in favor of clients who, on paper, would have suffered even more if the algorithm did not act otherwise. Here the algorithm must be able to distinguish between clients who really need the money, and who show good faith and commitment to repaying the loan, from those clients who are out to get cheap money without any intention of repayment. In this case, there are many specific details involved; the idea I am proposing is only that the algorithm should follow the Buddhist principle of trying to relieve suffering as best it can, based on the information available to it at the time.

Some may object to this proposal, saying that giving AI its own discretion in making more ethical decisions will inhibit the freedoms of the human user in applying AI in any way he or she sees fit. Furthermore, there is no guarantee that the algorithm will act as ethically as intended. These are legitimate concerns. Nonetheless, installing a component that inhibits the user from performing certain actions is not a new principle. For instance, some cars will not start unless the driver is wearing a seatbelt. The AI that refuses to follow certain orders from the user acts in the same way. Such a car limits the freedom of the user, but this is still seen as a strong safety feature. Additionally, how do we know that the AI, when given this amount of freedom, will always act ethically? For the artificial general intelligence (AGI) of the future, this is a serious matter because AGI's are capable of thinking on their own. Therefore, it is in our best interest to guide its development towards being both intelligent and ethical. For today's more specialized AI, however, safety devices should be installed or programmed so that the algorithm functions to promote ethical action.

In fact, giving AI the ability to act ethically is possible with today's technology. This does not necessarily mean that the AI is endowed with consciousness and free will. Instead, the AI is equipped with algorithms to act ethically and compassionately from the beginning. The microloan software will act ethically if it takes the interest of its clients into account. This might not maximize the bank's profit, but the social cost of being inflexible when loan decisions are analyzed and approved could be greater. As an increasing number of loan decisions are made autonomously by algorithms, having an ethical algorithm seems essential.

# Objections and Replies

During a series of meetings held by the Association of Pacific Rim Universities (APRU) under the project titled "AI for Social Good", my proposal benefited from a number of comments and helpful criticisms from my colleagues, who challenged me to develop a better, more defensible position on this topic. The first objection focused on the claim that ethics should be encoded into the algorithm or the inner programming core of the AI software. The objection is that it makes ethics too narrow and technical. According to this objection, ethics coding would result in the AI system being estranged from its social, cultural, and economic environment, leading to the system not being relevant to the aims of the forum. First, it should be noted that I would not advocate that social and cultural considerations should be taken away from ethical deliberation. This is just not possible, because ethics is always naturally embedded in the set of practices that surround any technical product, which is something that has been recognized by technology philosophers for a long time. For example, a car that remains stationary until the driver puts on a seatbelt, is an example of encoded ethics. According to my analysis, a car that neglects to warn the driver to wear a seatbelt and does not take appropriate action to ensure that he or she does so, is unethical. In the same vein, it is also ethical for microloan software to take more data points than required to ensure that loans are repayable. Sure enough, a program that makes an accurate risk calculation for a loan would to some extent be an ethical program. However, if this is all the software does, then its degree of ethicality is limited. It needs to consider other factors too, such as the condition of the loan applicant (e.g., economic status, children, health, etc.). It would be more prudent for the program to provide a loan under certain economic conditions, such as the current COVID-19 pandemic. The act of coding ethics into the inner workings of an AI program does not imply that the coder and employer are isolated from the surrounding socio-economic conditions or social environment. On the contrary, it shows that the coder and tech company value ethics, and must pay close attention to the needs and values of the society in which they intend to use the software.

The second objection builds upon the final sentence in the paragraph above. When a programmer encodes ethics into a machine, who will ensure that these ethics are correct? In other words, who or what would guarantee that the programmer does not put forward their own personal agenda and values into the software? In order to answer this question, one has to bear in mind that the programmer cannot, in fact, neglect the needs and values of society. If the programmer neglects those values and injects his or her own personal beliefs into the machine, it is likely that the machine would act strangely and be unusable. Software containing an idiosyncratic set of values would be condemned by users and thus would not be successful. The manufacturer would also have a strong interest in ensuring that the consumer receives a desirable service. Hence, the software would need to be tested repeatedly, not only for safety and quality control, but also for ethical quality.

According to the third objection, coding ethics into a machine is too narrow; the program must learn its ethics by interacting with its environment. Instead of taking all the cues from the programmer, an intelligent AI should be able to learn what is right and wrong from its interaction with other people. The more people it interacts with, the better it becomes at learning right and wrong. This is just like how a child learns ethics—to live in a social environment with parents, siblings, friends, and so on. There is just no way for an algorithm to understand ethics through code alone. This is a valid objection, but the coding is only a part of the larger program, which involves teaching a machine to be compassionate. Since we do not have AGI level machines yet, we have to see how specialized, blind ASIs (artificial specialized intelligence) can exhibit behaviors that we deem to be (approximately) compassionate. At this stage, we would be glad if AI could deliver social good, even without being conscious. The AI could be encoded in such a way that it knows how to learn ethical principles. Humans are already hardwired to become ethical, since altruism and cooperation among members of our species has been fundamental throughout our evolution. After

all, understanding ethical and social cues would be a very strong achievement for AI, but would still require coding for this possibility to occur.

The final objection explains that coding ethics into a machine implies that programmers and software companies do not care for, and are not accountable to, society at large. Again, this does not have to be the case. There is no logical link between coding ethics into an algorithm and the programmer and employer being unaccountable to society. We have seen earlier that the programmer and software company must ensure that their products meet the requirements set by consumers and society; furthermore, they are still a part of society and need to follow specific laws and regulations.

The objections and comments from my colleagues largely focus on the relation of coding to its socio-economic context. This is an important matter, and in conclusion I would like to argue that coding must be embedded within its contexts. More specifically, this means that coding must only be one aspect of the overall systematic practice of ensuring that AI is ethical. Nevertheless, without an emphasis on coding, there is no definitive way in which the design of AI could directly contribute to a better society. For this to happen, the components of an ethical AI need to be translated into a language that a computer would understand. That is, the ethical components need to be made operationalizable, and they need to be pared down into basic steps for a computer to follow. Most importantly, the ethical vision must be clear, and the operationalization needs to adhere to it closely.

## Conclusion and Recommendations

I would like to end this paper with a number of recommendations, both to public and private sectors, so that an ethical AI for social good can be fully developed and deployed. The recommendations are as follows:
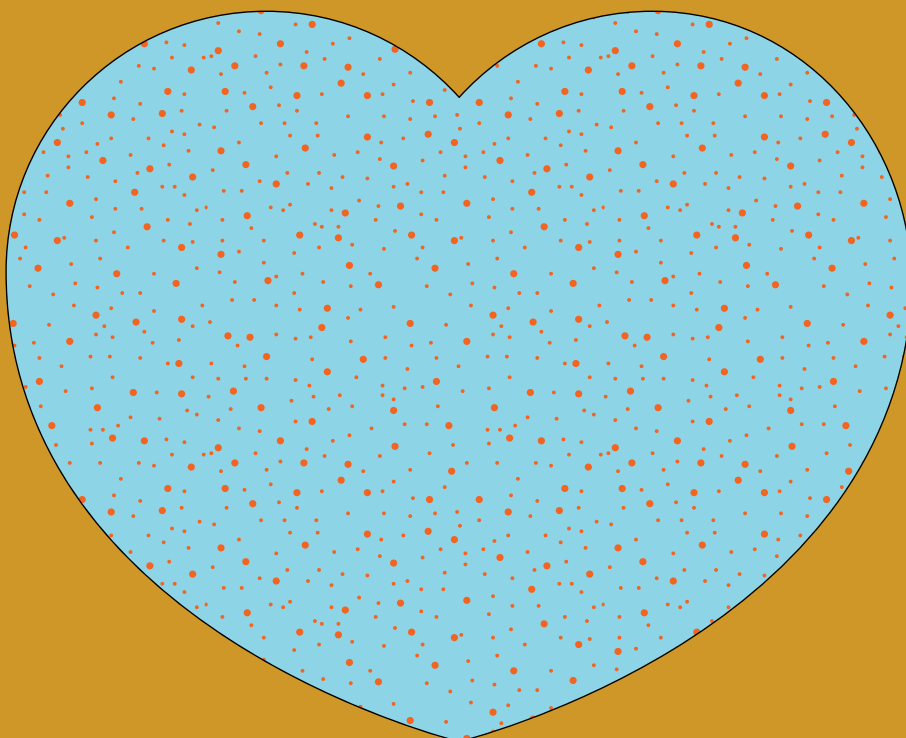
**Recommendation 1:** Programmers and software companies must implement compassionate AI programs, which is the key message of this article. No matter what kind of "social good" the AI is supposed to bring about, the software needs to be compassionate and ethical in the Buddhist sense. I have specified in some detail as to what being compassionate for AI actually means. Basically, the AI needs to realize that all things are dependent on all others (interdependence) and that the AI needs to show actual commitment to improving the condition of everyone in society (altruism). In order to make this recommendation feasible, the components of compassion need to be translated into algorithmic steps for the computer. In other words, the software needs to be coded in such a way that it becomes ethical. However, the coding must not be alienated from its socio-economic and historical contexts. That is, the software companies responsible for manufacturing AI programs must function as responsible and contributing members to society. No matter what kind of social good the AI is intended to bring about, this is a necessary requirement. The paper has shown that some applications that are being developed in the AI for Social Good program, such as flood forecasting, can indeed be used for nefarious purposes. This can happen when the information gained from the AI is used to gain unfair personal advantages. There should be ways within the design and programming of AI itself to prevent this, insofar as it is technically feasible. Abuses of flood forecasting information is an example of how the work of AI, which may originate from good intention, can be used in such a way that the AI itself becomes a culprit in an unethical action, such as hoarding or implementing flood prevention programs that privilege certain groups over others. Software companies need to be aware of this possibility and take the necessary steps to prevent it from happening.

**Recommendation 2:** The public sector needs to ensure that rules and regulations are in place in order to create an environment that facilitates the development of ethical AI for social good. Such rules and regulations will ensure not only that private companies have a clear set of directives to follow, but also public trust in the works of the private sector (assuming the work of creating AI software belongs to the private sector). Furthermore, even in a situation where the development of AI falls largely on the public sector, such as in Thailand, where the private sector is still rather weak in original research and development, the rules are also applicable. For example, the rules could provide incentives for software manufacturers to be more ethical. It needs to be made clear to all parties that there are material benefits to being more ethical. The belief that becoming ethical runs counter to profit maximization is shown to be unfounded. Realizing the objective of a private company must be embedded in the context of consumer trust; without the latter, it is hard to imagine how this type company could flourish in the long run.

These two recommendations make it clear that AI will create social good that truly answers people's needs and suffering. AI in the future may, or may not, become conscious and attain the level of superintelligence in the sense advocated by Nick Bostrom (Bostrom, 2014). In any case, AI needs to be made ethical at this time, as there is a decreasing window of opportunity to do so.

## Acknowledgments

# References

Berendt, B. (2019). AI for the common good?! pitfalls, challenges, and ethics pen-testing. *Paladyn, Journal of Behavioral Robotics*, 10(1), 44-65.

Bostrom, N. (2014) Superintelligence: *Paths, Dangers, Strategies*. Oxford University Press.

Cowls, J., & Floridi, L. (2018). Prolegomena to a white paper on an ethical framework for a good AI society. https://ssrn.com/abstract=3198732 or http://dx.doi.org/10.2139/ssrn.3198732

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People— An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707.

Gal, D., Perspectives and Approaches in AI Ethics: East Asia (June 7, 2019). Dubber, Markus, Pasquale, Frank, and Das, Sunit, (Eds.) *Oxford Handbook of Ethics of Artificial Intelligence, Oxford University Press*, Forthcoming. https://ssrn.com/abstract=3400816 or http://dx.doi.org/10.2139/ssrn.3400816

Green, B. (2019). "Good" is not good enough. AI for Social Good Workshop, NeurIPS (2019). https://www.benzevgreen.com/wp-content/uploads/2019/11/19-ai4sg.pdf

Hagendorff, T. (2019). The ethics of AI ethics: an evaluation of guidelines. *Arxiv.org*. https://arxiv.org/abs/1903.03425

Hongladarom, S. (2004). Growing science in Thai soil: culture and development of scientific and technological capabilities in Thailand. *Science, Technology and Society*, 9(1), 51-73.

Hongladarom, S. (2020). *The Ethics of AI and Robotics: A Buddhist Viewpoint*. Rowman & Littlefield.

Hughes, J. (2012). Compassionate AI and selfless robots: a Buddhist approach. In Patrick Lin, Keith Abney, and George A. Bekey (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge*, MA: MIT Press.

Kearns, M., & Roth, A. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.

Mittal, V. (2017). Top 15 Deep Learning applications that will rule the world in 2018 and beyond. *Medium.com.* https://medium.com/breathe-publication/top-15-deep-learning-applications-that-will-rule-the-world-in-2018-and-beyond-7c6130c43b01

Taddeo, M., & Floridi, L. How AI can be a force for good. *Science* 361 (6404), 751-752. DOI: 10.1126/science.aat5991.

The European Commission's High-Level Expert Group on Artificial Intelligence: Draft Ethics Guidelines for Trustworthy AI. (2018) Working Document for Stakeholders' Consultation. Brussels, 18 December 2018.
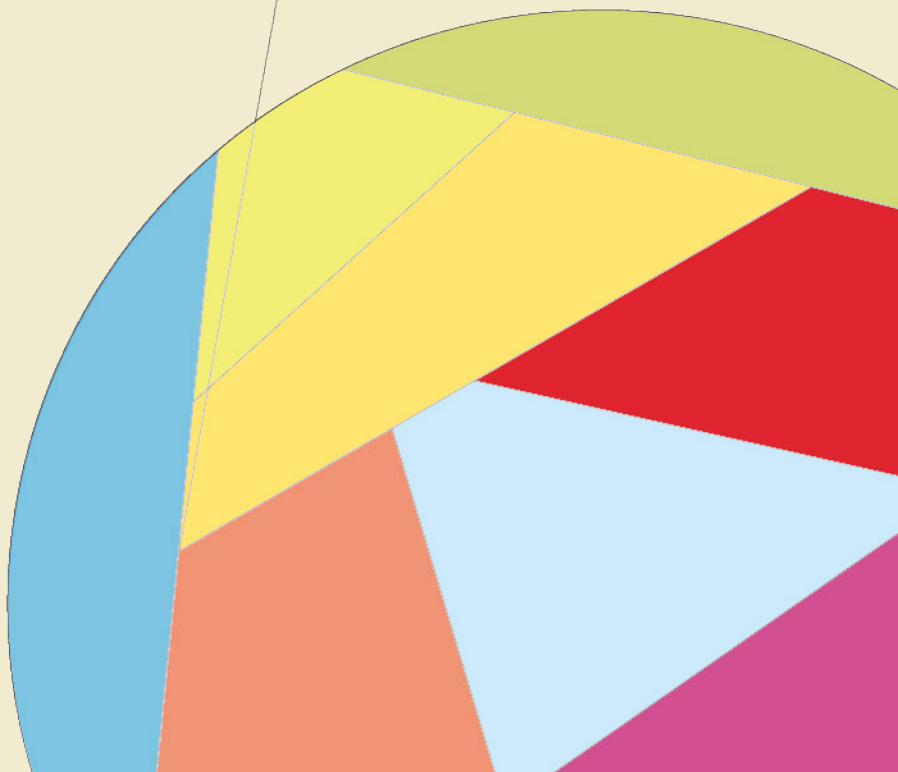
# Moralizing and Regulating Artificial Intelligence:

# Does Technology Uncertainty and Social Risk Tolerance Matter in Shaping Ethical Guidelines and Regulatory Frameworks?

**M. Jae Moon**
**Iljoo Park**
Institute for Future Government,
Yonsei University

# Introduction

Artificial intelligence (AI) is considered one of the most powerful developments in computer science, which affects every aspect and sector of society. While we are increasingly paying attention to its significance and impact, we do not yet know how and to what extent it affects the replacement and creation of jobs, industrial transformation, and lifestyle changes, which causes uncertainties and risks related to AI. Due to these underlying uncertainties and risks, there has been a growing demand for regulating and moralizing AI in order to minimize AI-caused uncertainties and risks. It is hoped that AI regulation will help to sustain its positive impact on society as a whole. With growing social fears and uncertainties, there has been increasing demand for a specific and proactive approach towards dealing with AI. Responding to these demands, governments and key international actors have attempted to provide regulatory frameworks and ethical guidelines for this rapidly developing technology. This study aims to review the uncertainty and risk issues of disruptive technologies such as AI, and assess their socio-economic and political impacts on society. This study will also discuss how key stakeholders (i.e., governments, industries, international organizations, NGOs, etc.) craft ethical guidelines/principles as well as review how different countries establish AI regulatory frameworks, particularly for autonomous vehicles (AVs).

Tzur (2017) argues that technological advancements fundamentally change the paradigm of regulatory mechanisms, while a conventional regulatory political framework (Wilson, 1980) seems to fail to offer an effective explanation for the nature of emerging disruptive technologies (i.e., AI, gene editing, blockchain, etc.), simply because defining who should benefit and who should bear the costs is quite uncertain and dynamic. Because of uncertainties regarding cost-benefit distributions as well as the opportunities and risks of emerging disruptive technologies, many countries appear to have adopted differing regulatory approaches to these technologies. For instance, national regulatory positions vary widely among different countries regarding the acceptance of cryptocurrencies (i.e., Bitcoins) as legal tender and the banning, regulation, or encouragement of cryptocurrency

exchanges. Notably, some countries such as Japan and the US have relatively light regulatory positions towards cryptocurrencies, while others including China and Korea have very restrictive policies. Likewise, regulations of disruptive technologies also differ in content and intensity from country to country. While some governments are in a strict regulatory position, others remain in an active deregulatory position by introducing regulatory sandboxes. Furthermore, the uncertainty and new forms of risk posed by these technologies (Slovic, 1987) demand social, industrial, and often international agreement, as well as discussion on ethical requirements and technological standards to ensure the maximization of social benefits and the minimization of social risks of these disruptive technologies.

In general, governments enact regulations to correct market failures, pursue collective and public interest goals, and to prevent potential social problems caused by the excessive pursuit of private interests. However, individual regulations do not always meet public expectations or help achieve intended social goals. Regulatory decisions on disruptive technologies are often not timely, primarily because of the lag between the emergence of technology-driven social issues and regulatory policy decision-making. Views regarding the regulation of novel technologies also often vary widely because of country-specific contextual factors—including legal systems, influence of various interest groups, and the ethical perspectives of the general public, which determines the social risk perceptions of the public.

This study uses cross-country comparative case studies by examining the similarities and differences of regulatory actions caused by levels of certainty, as well as the tolerance of social risks for technologies in given countries. As an example, this study will examine regulatory approaches to AVs, which is a product of AI and robotics technology. We will examine the

US and three Asian countries, namely China, Japan, and Korea. The aforementioned Asian countries are major economic players in the region, and are all interested in disruptive technologies for the potential implications of economic and social development. The US has been included as a base for comparison since it is more market-oriented than other countries, while the three Asian countries are somehow paternalistic.

Due to the disruptive nature of emerging technologies such as AI and related technologies including robots, AVs, drones, etc., there is no particular consensus regarding how disruptive technologies should be regulated and moralized through social interests in those technologies, as well as research interests in the intertwined relationship between technological advancements and regulations. Despite growing interest in disruptive technologies and related ethical guidelines and regulations, limited research has been conducted in this field. In particular, a comparative analysis of ethical guidelines for AI and different national responses to disruptive technologies have been somewhat lacking, primarily because there is no clear measure of regulatory stringency as the basis for comparative studies of regulation politics (Brunel & Levinson, 2013). In order to fill this research gap, this study aims to look at key ethical elements of AI, and then determine how and why countries develop different regulatory approaches to the same technologies.

Along with the growing interests in AI, governments, research institutes, international organizations, and industries initially began to pay attention to ethical frameworks for AI, as many are puzzled about the potential consequences and ethical dilemmas. For example, an ethical dilemma on this subject is how an autonomous vehicle should deal with an unavoidable accident, where the car must decide whether to kill an innocent bystander or the five passengers inside the vehicle. It is also imperative to question who

should be responsible for an incident involving an autonomous vehicle, among AI programmers, vehicle manufacturers, vehicle sellers, drivers, and others.

As proposed by a group of experts on AI commissioned by the OECD, an ethical guideline specifies and addresses core values in developing, manufacturing, and using AI and AI-loaded machines. In fact, it will not be long before ethical guidelines and principles for AI are offered by governments, international organizations, private companies, and NGOs. Reviewing 84 documents of ethical principles and guidelines, Jobin et.al. (2019) found that most of these documents (88%) were released after 2016 by private companies (22.6%) and government agencies (21.4%).

We will first discuss technology uncertainty and social risk in the context of disruptive technologies. Then, we will review the development of ethical guidelines for AI developed by different actors as a loosely institutional effort to moralize AI technologies. Next, we specifically examine the different regulatory positions of four selected countries to AVs. Finally, policy implications are discussed and policy recommendations are presented.
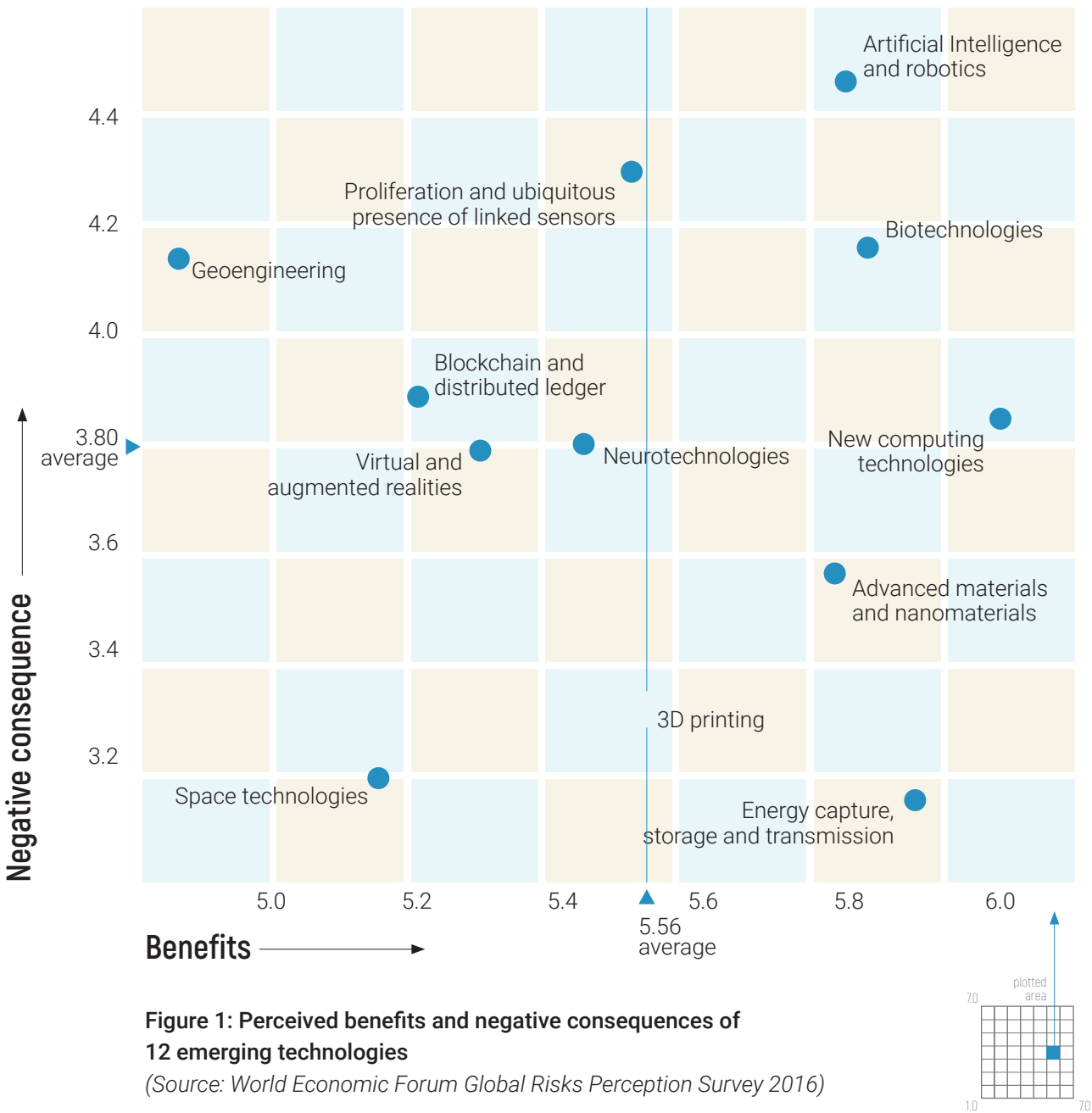
## Determinants of Regulating and Moralizing Disruptive Technologies: Technology Uncertainty and Social Risk Tolerance

### Disruptive technologies: benefits and risks

Since being presented by the World Economic Forum in 2016, there has been a growing interest in disruptive technologies which are often proposed as technological engines for the fourth industrial revolution. Figure 1 shows the different levels of expected benefits and costs from each technology. The World Economic Forum (2016) surveyed professionals

in each country, asking about their perceptions of the benefits and negative consequences of 12 major emerging disruptive technologies. Participants perceived AI and robotics as the most beneficial and risky technologies, while they perceived blockchain technology as moderately beneficial and risky. Moreover, people tend to perceive both biotechnologies and neuro-technologies to be more beneficial and riskier than blockchain technology.

Despite variations in the perceived benefits and risks of those disruptive technologies, many stakeholders have raised their concerns over the potential risks of such technologies. As such, they have demanded for alternative ways of moderating and minimizing the risks, which often results in informal/unofficial forms of ethical principles and formal/official forms of regulation. While the former is presented as a set of soft, suggestive, and general principles, the latter is a set of hard, legally binding, and specific rules. The former is discussed and manufactured by various stakeholders of different sectors (private, non-profit, and public sectors) at different levels (i.e., local, national, and international), whereas the latter tends to be made by executive or legislative branches through formal rule-making and legislative processes, because each country makes its own regulatory decisions as technological risks and interest conflicts among stakeholders gradually mount. Recently, ethical standards and regulations have been discussed and proposed in the European Union (EU), the OECD, and other economic communities to moralize as well as control (regulate) technologies. While there is a general consensus in the nature and scope of ethical principles for AI, there is no consensus in regulatory frameworks among different countries. Moreover, the governmental regulatory decision can fall even farther behind when the potential costs and benefits of a technology are uncertain.

**Figure 1: Perceived benefits and negative consequences of
12 emerging technologies**
*(Source: World Economic Forum Global Risks Perception Survey 2016)*

## Regulatory lag and regulatory paternalism

Regulators are often uncertain as to whether or how to address the risks (World Bank, 2016). In particular, regulators are uncertain and unclear about assessing the potential benefits and risks of emerging technologies, which makes regulating disruptive technologies even more challenging than conventional technologies (Hunt and Mehta, 2013). Generally, regulations tend to be reactive rather than proactive, which often causes regulatory lag. While regulatory lag is partially a result of market-based and non-interventionistic policy position, it often causes tardy responses to previous problems that could have been addressed in advance.

On the contrary, regulatory paternalism also plays an important role in driving proactive regulations to minimize potential risks. Paternalism originally referred to the ideological belief that governments should intervene to protect people—similar to protecting their children. Thus, regulatory paternalism involves paternalistic regulatory action on the part of governments. Paternalism lies behind many regulatory measures beyond specific instances (e.g., seatbelt and safety helmet laws); it is also the driving force behind the prohibition or control of certain risk-generating products and services. In fact, citizens of contemporary risk-obsessed societies expect their governments to provide them with protection (Ogus, 2005). To overcome excessive regulations formulated by regulatory paternalism, some countries have recently adopted temporary deregulation schemes such as a regulatory sandbox, which is a testing ground that is protected against any possible regulation. This supports a flexible and lenient regulatory position to maximize potential economic and social benefits of various disruptive technologies.

## Determinants of cross-country regulation differences

Based on the "psychometric paradigm," Slovic, Fischhoff, and Lichtenstein (1982) conducted a classical study regarding the risk perception of people and offered a solid framework to understand the cross-country regulation difference on disruptive technologies. They suggest two significant factors to distinguish technologies: dreadfulness and unfamiliarity. Dreadfulness refers to the extent to which a technology can be controlled not to be catastrophic, which is understood as a measure for technological risk. Unfamiliarity refers to how much a technological risk is observable, which is considered as a technology uncertainty. It implies that subjective perception is an important factor to the classification of technologies besides objective criteria. It should be noted that these terms are not absolute, and instead used as relative terms. For instance, nuclear power can be a more dreadful and less unknown technology than dynamite, which is a less dreadful but more known technology.

While "uncertainty" and "social risk" are considered to be independent, they are somewhat related since technology uncertainty often causes a higher level of social risk of a particular technology in a society. As a result, the social tolerance of a particular risk would be a significant factor in a country since the response to one technology would be different for other countries, although the objective technological risk would be identical. This leads to specific regulatory positions for different technologies because certain countries may want to control the potential technological risk and take various regulatory measures (e.g., law enactments) to restrict the reckless research, development, and utilization of technology.

## 1. Technology uncertainty

Technological "unfamiliarity" (Slovic et.al., 1982) is somewhat similar to technology "uncertainty", though the term "uncertainty" may not be used in a strictly defined sense since it is commonly used by many people in different senses (Downey and Slocum, 1975) or often poorly understood (Fleming, 2001). Despite this poor understanding of "uncertainty", it is generally accepted that the degree of technology uncertainty may vary depending on controllability, which is directly related to the level of safety and potential risk of a particular technology. According to Milliken (1987), the three common definitions derived from psychology and economics for "uncertainty" are (1) "an inability to assign probabilities as to the likelihood of future events", (2) "a lack of information about cause-effect relationship", and (3) "an inability to predict accurately what the outcomes of a decision might be". Similarly, we can define technology uncertainty as "the inability to measure the likelihood of a future event and the outcome with probabilistic function and to infer the causal outcome made by a particular disruptive technology".

We argue that *uncertainty about the spillover effects from technologies themselves results in* cross-country variation in regulatory decisions on disruptive technology. For example, the difficulty of predicting the costs and benefits of a technology causes regulatory lag since this can obstruct timely regulations. Governments are likely to identify disruptive technologies based on the extent to which the expected costs and benefits are easily measured. If the costs and benefits derived from a technology can be predicted quickly, the regulatory policies can be developed more promptly. Otherwise, governments may postpone strict regulatory decisions if a technology has the potential to cause harm in ways that cannot be foreseen during the innovation process, preventing them from quickly predicting the costs and benefits the technology could generate. We define such technologies as "uncertain technologies". It should be noted that the regulation of uncertain technologies is also affected by the degree of uncertainty that a particular society should and can tolerate (Kolacz et al., 2019).

In contrast to the uncertainty of expected outcomes from any given technology, *responsiveness to the global consensus* is a significant factor for converging similar regulatory positions. Although it may be challenging to make a public consensus between scientists and the general public (Kahan, Jenkins-Smith, & Braman, 2011), the existing consensus or standards can apply to regulatory decisions regarding emerging technologies. Recently, a global consensus led by international and regional organizations such as the EU, the OECD, and the WHO has also been made, which shapes the nature of regulatory positions of countries that are not necessarily obligated to follow the global standard (Kerwer, 2005).

## 2. Social risk tolerance

Another reason for differences in regulatory responses between countries is that some countries have different levels of tolerance for social risks. Uncertainty of one technology makes people eager to prepare for potential risks or hazards. We focus on the fact that the preparation level for an uncertain technology can differ depending on the country. Social risk tolerance is closely related to uncertainty avoidance; people who prioritize avoiding uncertainty are likely to control uncertain situations by imposing strong schemes such as regulations. Empirical studies in various areas — e.g., Kanagaretnam et al. (2011) — examine the relationship between high risk perception and low uncertainty avoidance.

Hofstede's 6-D model of national culture is considered one of the major measurements of the general public's uncertainty avoidance. It attempts to measure the degree to which members of a society feel uncomfortable with uncertainty and ambiguity (Hofstede, 2015). According to Hofstede's score out of 100, Japan (92) and Korea (85) have somewhat higher uncertainty avoidance than China (30) and the US (46). Note that the interpretation of this index has been made cautiously because Hofstede originally developed his theory from a management perspective to recognize the difference between diverse cultures. That said, it helps to draw a better understanding of the cultural differences among countries in many aspects, such as uncertainty avoidance. Uncertainty avoidance is different to risk avoidance, but is related

to anxiety and distrust towards the unknown (and vice versa), with the desire to have fixed practices and rituals as well as understanding reality (Hofstede, 2015).

Exploring the determinants of social risk tolerance levels could provide substantial insight into cross-country differences in regulatory decisions regarding disruptive technologies; however, discussion of such an approach in prior research is scarce. We identify the following three main factors that define countries' different tolerance of social risk: (1) *legal traditions and the efficiency of legally challenging regulations*, (2) *competition among interest groups*, and (3) *ethical concerns*.

First, *legal traditions and efficiency of legally challenging regulations* can generate differences in regulatory decisions among countries. Numerous studies, including Beck et al. (2002) and Hail and Leuz (2006), examine the relationship between countries' legal origins and levels of economic development, finding the nations' legal origins significantly impact their financial development. In particular, Beck et al. (2002) suggests that differences in countries' legal origins help explain differences in their levels of financial development.

Furthermore, some empirical studies have identified differences between common law and civil law countries in terms of regulation decisions. For instance, Djankov et al. (2002) finds that, at comparable levels of development, French civil law countries tend to have heavier regulations, less secure property rights, and fewer political freedoms than common law countries. Moreover, Charron et al. (2012) also mention that countries' legal origins could explain cross-country differences in judicial independence and government regulations of economic life, which can be summarized as the quality of institutions, as well as low degrees of corruption and high degrees of the rule of law, which in essence are desirable social and economic outcomes. They suggest that because of stronger legal protections for outside investors and less state intervention, countries with a common law tradition have achieved higher economic prosperity

and quality life than civil law countries. La Porta et al. (2008) even summarize their series of articles (La Porta et al. 1997, 1998, 1999) to address the prevalent impact of a wide range of desirable organizations and social outcomes of nations' legal traditions and other related articles to develop a so-called "Legal Origins Theory" (Charron et al. 2012).

*Competition among interest groups* can also generate differences in countries' regulatory decisions. Gai et al. (2019) explain that regulatory complexity is a consequence of lobbying. They focus on the fact that lobbyists may be able to persuade policymakers or politicians to give their interests to more favorable regulatory treatment, which leads to additional complexity and fragmentation across countries, especially when it comes to financial regulation. In addition to the appeals of individual groups, conflict among many interest groups can significantly affect countries' regulatory decisions. For instance, interest-group politics are heavily involved in cryptocurrency regulation; debates regarding the use of cryptocurrency worldwide is intense, and many stakeholders are involved in this discussion. According to Houben and Snyers (2018), numerous players are involved in the cryptocurrency debate and they all play particular roles: cryptocurrency users, miners, cryptocurrency exchanges, trading platforms, wallet providers, coin inventors, and coin offerors. In addition to these players, policymakers such as the International Monetary Fund (IMF), the Bank for International Settlements, and the World Bank have their own views on cryptocurrency. The groups who utilize cryptocurrency are expected to experience the associated benefits, costs, and discussions, which are still ongoing.

*Ethical concerns* can also lead to differences in countries' regulatory decisions. Such concerns may be related to general public safety or the religious views of various groups. In particular, regulations regarding genetically modified organisms (GMOs) are affected by the ethical perspectives of countries' citizens. Such perspectives can be affected by religious beliefs or the general views of human morality. Globus and Qimron (2018) investigate the regulations and

cultural perceptions of different countries regarding GMO approval. Their study found that regulatory and supervisory procedures for GM crops and the foods produced from these crops differ because governmental approaches represent the differing responses of citizens and scientific communities. These policies also reflect a variety of cultures, environmental conditions, political pressures, and the interests of different groups such as farmers, agricultural companies, and environmental activists or agencies.

To summarize, we suggest that the regulation of disruptive technology might vary as a result of technology uncertainty and social risk tolerance, and that several socio-economic factors may generate variation in uncertainty and risk tolerance. Two different approaches have been suggested: (1) moralizing technologies based on ethical standards and (2) regulating technologies based on legal mechanisms. The former refers to the efforts of various stakeholders to promote desirable status or conditions through codes of conduct or moral principles, which are often voluntary instead of mandatory. The latter refers to legal actions by governments to mandate and enforce particular actions, or to prohibit illegal actions which in many cases lead to penalty or punishment. In the next section, we examine the evolution of ethical principles for AI and then survey regulatory actions regarding three selected disruptive technologies that pose different degrees of risk in four developed countries.

| | Ethical Approach | Legal Approach |
|---|---|---|
| **Mechanism** | Ethical standards | Regulatory laws |
| **Actor(s)** | Various stakeholders | Government(s) |
| **Nature** | Voluntary; Broadly defined and widely applied | Mandatory; specifically defined and narrowly applied |
| **Consequences** | Moral blaming | Punishment or penalty |

Table 1: Comparison of ethical approach and legal approach
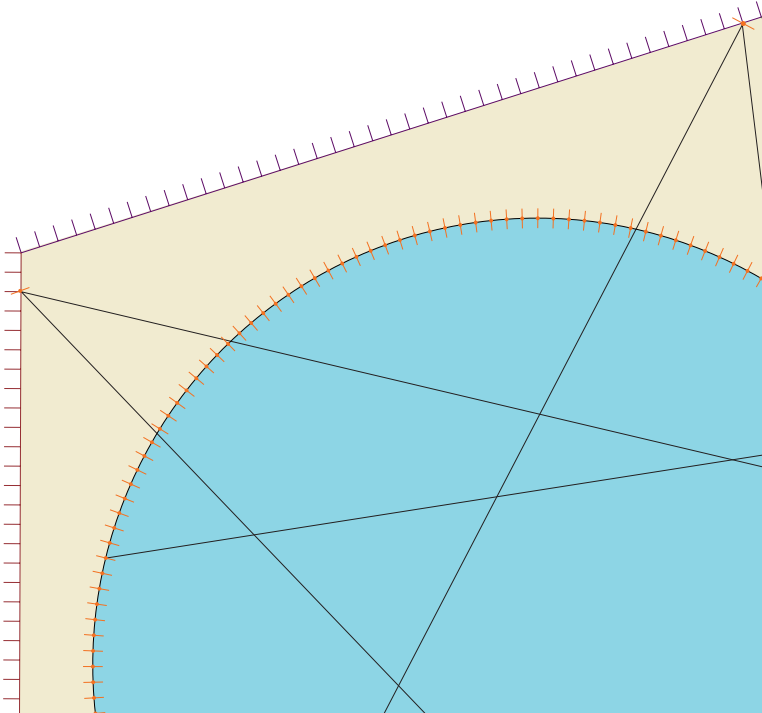
## Moralizing Disruptive Technologies: Ethical Guidelines and Principles for AI

Ethical AI (Jobin et.al., 2019), trustworthy AI (European Commission, 2018), and responsible AI (Microsoft, 2018) have been proposed and discussed among various stakeholders (e.g., academics, industries, governments, and international organizations), as AI was presented as a main driver for radical and disruptive changes (Jobin et.al., 2019). Although terms such as "ethical", "trustworthy", and "responsible" are used in documents that cover ethical guidance and principles, they all explain that we must handle AI in a lawful, ethical, and robust way throughout its entire lifecycle. Such guidelines include design, development, deployment, and usage (European Commission, 2018) by recognizing, preparing, and resolving the potential risks and negative impacts of AI in a society.

Ethical AI is often considered as a starting point for moderating any potential negative social and economic impacts of AI and AI-loaded devices, including automation and job replacements, intentional misuses and malevolent consequences, dissemination of social bias and its reinforcement, and an undermining of fairness (Jobin et.al., 2019). Reviewing and scoping 84 documents of ethical guidelines and principles, Jobin and her colleagues (2019) suggest that several key ethical principles are commonly identified including transparency, justice and fairness, non-maleficence, responsibility, and privacy. That said, there is no consensus on how these principles are interpreted and applied in the course of designing, developing, and using AI and AI-loaded devices.

Presenting trustworthy AI, the European Commission (2018) proposed three elements constituting trustworthiness including lawful AI, ethical AI, and robust AI. Lawful AI refers to the fact that AI should be bound by existing legal systems of local, national, regional, and international levels so that they bind any processes and activities involving the entire AI lifecycle. The European Commission (2018) suggests that lawful AI "should not be interpreted with reference to what cannot be done, but also with reference to what should be done and what may be done". In addition to legal compliance as a basic minimal requirement, ethical AI emphasizes the reference of ethical norms in particular because legal systems are often far behind and do not keep up with technological developments. Robust AI is presented to avoid or minimize the possible unintended negative consequences of AI in a society.

As shown in Figure 2, the European Commission (2018) suggests that all stakeholders including developers, deployers, and end-users should meet critical requirements for realizing trustworthy AI. Seven requirements are presented as follows: (1) human agency and oversight (fundamental rights, human agency, and human oversight); (2) technical robustness and safety (resilience to attack and security, fallback plan and general safety, accuracy, and reliability and reproducibility); (3) privacy and data governance (privacy and data protection, quality and integrity of data, and access to data); (4) transparency (traceability, explainability, and
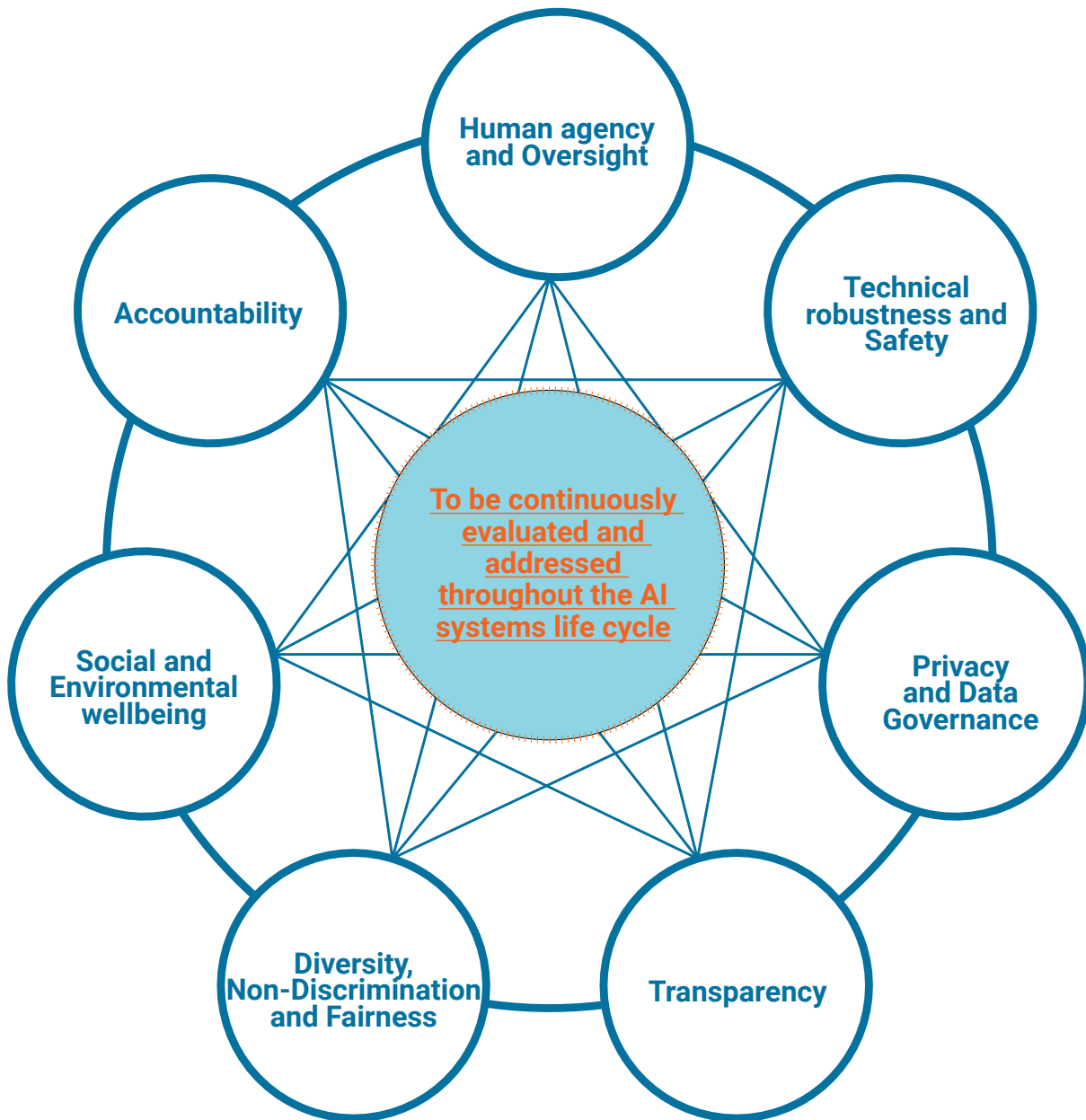
**Figure 2: Seven requirements for trustworthy AI and their interrelationship**
*(Source: European Commission (2018), p. 15.)*

communication); (5) diversity, non-discrimination and fairness (avoidance of unfair bias, accessibility and universal design and stakeholder participation); (6) societal and environmental wellbeing (sustainable and environmentally friendly, social impact, and society and democracy); and (7) accountability (auditability, minimization and reporting of negative impacts, trade-offs, and redress) (European Commission, 2018).

Similar to the Ethics Guidelines for Trustworthy AI by the European Commission, many organizations and governments have offered ethics guidelines and principles for AI. As summarized in Table 2, many documents have been formulated by private companies, government agencies, and academic institutions; many of which were formed in the US, UK, and EU institutions. Table 2 shows the breakdown of ethical guidelines and principles for AI by type, geographical location, and target audience.

| Type and Geographical Location | Classifications |
|---|---|
| **Type of Issuing Organizations*** | 19 private companies (22.6%), 18 government agencies (21.4%), 9 academic and research institutions (10.7%), 8 inter-governmental or supra-national organizations (9.5%), 7 non-profit organizations and professional associations (8.3%), 4 private sector alliances (4.8%), 1 research alliance (1.2%), 1 scientific foundation (1.2%), 1 federation of worker unions, 1 political party, 4 others |
| **Geographical Location of Issuing Organizations**** | 20 USA (23.8%), 16 international organizations, 14 UK (16.7%), 6 EU institutions, 4 Japan, 3 Germany, 3 France, 3 Finland, 2 Netherlands, 1 Iceland, 1 India, 1 Singapore, 1 Norway, 1 South Korea, 1 Spain, 1 UAE, 1 Australia, 1 Canada |
| **Target Audience***** | 27 for multiple stakeholder groups (32.1%), 24 for own employees of companies (self-directed) (28.6%), 10 for the public sector (11.9%), 5 for the private sector (6.0%), 3 for developers or designers (3.6%), 1 for organizations, 1 for researchers |

Source: Compiled by author from Jobin et.al. (2019).
* 4 documents are double counted and 4 are not classified
** 3 are not classified
*** 13 not classified.

**Table 2: Ethical guidelines and principles by type and geographical location**

Based on content analysis, Jobin and her colleagues identified 11 key ethical principles along with related values. Some key findings on ethical principles from the content analysis by Jobin and her colleagues (2019) are summarized in the following table. As the table indicates, transparency and related values (73/84) appeared the most, followed by justice/ fairness (68/84), among 11 key ethical principles including transparency, justice/fairness, non-maleficence, responsibility, privacy, beneficence, freedom/autonomy, trust, sustainability, dignity, and solidarity. Non-maleficence and responsibility are also primary principles which are found in 60 out of 84 documents.

| Ethical Principles | No. of Documents | Related Values |
|---|---|---|
| Transparency | 73 | Explainability, explicability, understandability, interpretability, communication, disclosure, showing |
| Justice/fairness | 68 | Consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution |
| Non-maleficence | 60 | Security, safety, harm, protection, precaution, prevention, integrity, (bodily or mental), non-subversion |
| Responsibility | 60 | Accountability, liability, acting with integrity |
| Privacy | 47 | Personal or private information |
| Beneficence | 41 | Benefits, well-being, peace, social good, common good |
| Freedom/autonomy | 34 | Freedom, autonomy, consent, choice, self-determination, liberty, empowerment |
| Trust | 28 | |
| Sustainability | 14 | Environment (nature), energy, resources |
| Dignity | 13 | |
| Solidarity | 6 | Social security, cohesion |

**Table 3: Ethical principles and related values**
*(Source: Jobin et.al. (2019), p. 7.)*

As noted in the earlier section, international organizations such as the EU have been actively working on formulating ethical guidelines for AI. For example, the European Parliament took an initial action by asking the European Commission to assess AI's social impacts, which led to a set of "recommendations on civil law rules on robotics" in early 2017 (Madiega, 2019). This was followed by the Commission's coordinated plan on AI for EU member countries, which was later endorsed by the EU Council and then became a foundation for the Commission's Ethics Guidelines for Trustworthy AI (Madiega, 2019). The guideline formulated by the High-Level Expert Group on AI of the Commission is considered one of the most comprehensive frameworks for offering critical principles that various stakeholders should consider in designing, developing, and deploying AI. In particular, the guideline emphasizes the core nature of a "human-centric approach", which has been widely accepted beyond the EU. The nature of this human-centric approach to AI is summarized as follows:

> *The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed and deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoys a unique and inalienable moral status. This also entails consideration of the natural environment and of other living beings that are part of the human ecosystem, as well as a sustainable approach enabling the flourishing of future generations to come.[1]*

Emphasizing the lawfulness, ethics, and robustness of a trustworthy AI system from a lifecycle perspective, the guideline essentially promotes ethical principles for ensuring reliable and trustworthy AI. The guideline emphasizes seven key requirements for EU member countries including (1) human agency and oversight, (2) robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental well-being, and (7) accountability (Madiega, 2019).

## Regulating AI: The Case of Autonomous Vehicles in Different Countries

As noted earlier, regulatory instruments and levels of regulation vary widely from country to country. We conduct an exploratory comparison of the regulatory approaches of four major countries—China, Japan, Korea, and the US—in terms of the regulatory intensity of AVs. The three Asian countries were selected because they are considered as economic leaders, while also representing countries at different levels of economic development in the region. The US was selected as a basis for comparison, as the country represents market-based and relatively non-interventionist regulation policies.

### 1. Current status of autonomous vehicle technology development

An autonomous vehicle (AV) is a vehicle that can navigate by itself without human intervention (Taeihagh & Lim, 2019). According to SAE International (originally the Society of Automotive Engineers), automated driving can be divided into six levels, from 0 to 5 (the higher the level, the more automated the vehicle), based on the level of sophistication and automation. As Figure 3 summarizes, AVs are equipped with various autonomous features for driver supporting systems ranging from automatic emergency breaking (Level 0) to lane centering systems (Level 2: partial "hands off" automation), while "automated driving systems" also range from traffic jam "chauffeurs" (Level 3: conditional "eyes off" automation) to the highest level of complete driverless taxis in all conditions (Level 5: full "steering wheel" automation) (QVRTZ, 2019). Several carmakers, including Waymo, are already using level 4 AVs in some areas for ride-sharing or delivery services, but these vehicles have not yet entered the retail market. It has been said that substantive impact of AVs might begin when driverless automobiles are introduced in local areas.

---

1. Glossary section of the Ethics Guidelines for Trustworthy of AI (2019). Requoted from Madiega (2019), p. 3.
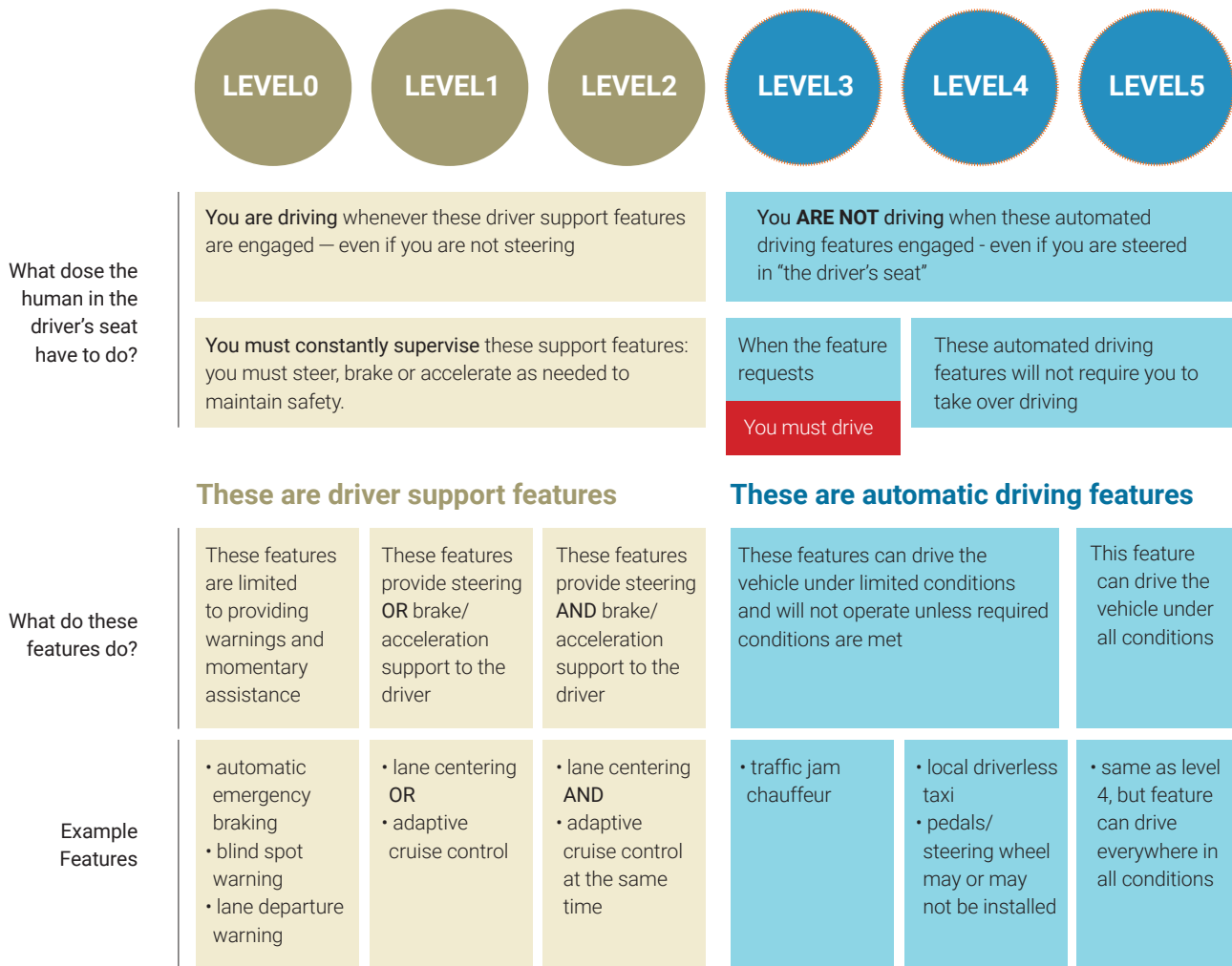
**Figure 3: Levels of autonomous vehicles**
*(Source: SAE (2018). https://www.sae.org/news/press-room/2018/12/sae-international-releas-es-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-stand-ard-for-self-driving-vehicles)*

## 2. Regulating autonomous vehicle

Table 4 presents a cross-country comparison of the specific regulations for AVs, particularly focusing on AV driving in China, Japan, Korea, and the US. We consider four regulatory issues: (1) whether the government permits autonomous driving, (2) whether the enforcement is legally binding, (3) whether the government can hold people liable based on laws or guidelines, and (4) whether the government provides any guidelines for users. We will not discuss license issues, since it has been debated at national levels. It should also be noted that no global consensus currently exists and nation states generally have strict requirements for drivers.

The three Asian countries under examination have prohibited autonomous driving when the driving is not for testing, and enforcement is legally binding. The US, however, has placed no strict restraints on autonomous driving; a bill that would establish the federal government's role in ensuring the safety of highly automated vehicles has been referred to a federal committee. All countries except China can hold persons (rather than AVs) liable based on these laws or guidelines; as it stands, China has no official guidelines regarding the issue. Furthermore, people who want to take autonomous driving tests

| | Prohibiting free autonomous driving itself | Legally binding enforcement | Holding persons liable based on the laws or guidelines | Offering guidelines for users |
|---|---|---|---|---|
| **China** | Yes | Yes | No | Yes |
| **Japan** | Yes | Yes | Yes | No |
| **Korea** | Yes | Yes | Yes | No |
| **US** | No (No strict restraint) | No (Referred the bill to the Committee) | Yes (Those who want to test AVs should obtain the state-designated insurance) | Yes |

**Table 4: The status of autonomous vehicle driving regulations (as of August 2019)**

must obtain state-designated insurance in Korea. Governments' provision of user guidelines for autonomous driving demonstrates their interests in the development of autonomous driving technology and commercialization. China and the US have user guidelines while Japan and Korea do not.

The US Congress passed a bill titled the *Safely Ensuring Lives Future Deployment and Research in Vehicle Evolution Act* (more commonly known as the "SELF DRIVE Act") in 2017. Proponents of the bill claim that by encouraging the testing and deployment of AVs, the bill establishes a federal role in ensuring the safety of highly automated vehicles. It has been received in the Senate, read twice, and referred to the Committee on Commerce, Science, and Transportation (The US Congressional Research Service, 2019). In addition to this bill, the US is the first country to introduce legislation to permit the testing of automated vehicles (UK Department for Transport, 2015). It has also introduced "A Vision for Safety 2.0," federal guidelines for the automobile industry and individual states regarding automated driving systems (ADSs) that builds on the National Highway Traffic Safety Administration's 2016 guidelines. This

document has two sections—voluntary guidance and technical assistance for states. The new guidelines focus on Levels 3 to 5 of the SAE International's automation classification, stipulating that entities do not need to wait to test or deploy an ADS, revising the elements of safety self-assessments, aligning federal guidelines with the latest developments and terminology, and clarifying the role of the federal and state governments. The guidelines emphasize their voluntary nature and do not include with compliance requirements or enforcement mechanisms. They represent an attempt to establish best practices for state legislatures, outlining the common safety-related components of ADSs that states should consider incorporating into their legislation. Additionally, they include the US Department of Transportation's view regarding federal and state roles and offers best practices for highway safety officials.

China is also preparing regulations to ensure safe AV testing. Notably, Chinese regulations and policies regarding autonomous driving are seen as relatively moderate compared to their strict control of some other aspects of driving, such as restrictions stating that public maps can only be accurate to a scale of

50 meters at most, and that drivers must keep both hands on the steering wheel at all times (KPMG International, 2018). The road-testing regulation was established in April 2018 and the guidelines for building safe, closed test sites were released in July 2018 (Xinying, 2019). The Chinese do not appear to be very concerned with safety and liability issues; their concerns focus on the technological availability of AVs and economic consideration related to their use (Dickinson, 2018).

Likewise, Japan is preparing the commercialization of level 3 AVs and will enact a new legal amendment for autonomous driving. The National Diet of Japan passed a bill amending the current Road Transport Vehicle Act to include "automatic operating devices" as a vehicle in May 2019. In addition, it passed another bill that allows people to use level 3 AVs in certain conditions and to use cell phones during autonomous driving (Matsuda et al., 2019). Although there has been some progress in AV-related regulations thanks to the May 2019 amendments of Japan's Road Traffic Act, Matsuda and his colleagues (as quoted below) stressed that there are still several issues to be resolved in future.

> *"… One of the main outstanding issues is determination of the rules for criminal and civil liabilities in the event of a traffic accidents involving self-driving vehicles. Because these provisions have not yet been updated, a driver may still be held responsible for criminal or civil liabilities for a traffic accident caused by a vehicle under automated driving even if the driver operated the self-driving vehicle properly. This issue affects not only drivers but also manufactures and insurance companies, and is therefore likely one of the thornier issues remaining to be resolved"* (Matsuda et al., 2019).

In Korea, the Road Traffic Act, Automobile Management Act, and Automobile Damages Guarantee Act currently regulate the use of automobiles, but that will change in 2020 when the Act on the Promotion and Support of the Commercialization of Self-driving Cars comes into force. The Road Traffic Act regulates traffic problems and establishes rules for safe operation. It presumes the presence of a driver who is required to manipulate the steering wheel and braking system. However, the Automobile Control Act defines AVs as cars that can be operated without any driver or passenger input. The Enforcement Rules of the Act, enacted in 2016, specify the requirements for the safe operation and testing of AVs, meaning that the laws are in conflict with each other to some extent regarding whether "a driver" can refer to an automated system. At present, the Ministry of Land, Infrastructure, and Transport requires a temporary operation permit for the testing of AVs, and the "Requirements for Safe Operation of Autonomous Vehicles and Trial Operation Regulations (as of March 31, 2017)," stipulate that a preliminary test of 5,000 km must be conducted (Ministry of Science and ICT and KISTEP, 2018).

The KPMG International's annual reports provide insight into the current state of AV testing. The reports evaluate countries' AV readiness and AV testing restrictions, giving countries scores out of seven based on reviews of media articles, government press releases, and government regulations. A higher score indicates that the country's regulations support AV use and impose fewer restrictions on when, where, and how testing of AVs can occur (KPMG International, 2019). According to the report, among the four countries considered in this study, Japan has the strictest regulations on AV testing with a score of 0.333, while Korea and the US have somewhat fewer restrictions on AV testing, both receiving scores of 0.833; China's score was 0.5 in AV regulation (KPMG International, 2019). The scores of 2018 are largely the same, although a different scale was used (KPMG International, 2018).[2] Similar to AV regulation score, Korea and the US have higher scores than China and Japan in terms of institutional responsibility for AVs (KPMG International, 2019). According to the indicator of the AV-focused government agency by the KPMG International, South Korea's score is 0.857 and the US is 0.714. China's score of consumer AV acceptance is 0.643 and Japan is 0.571, which is the lowest among the four countries (KPMG International, 2019). Considering the fact that regulations are often affected and influenced by the voices of private businesses, the number of AV firms in a country might be a factor which is closely associated with the nature and level of regulations on AV test driving and safety. According

---

2. According to the 2018 scores on AV regulation, Japan, China, Korea, and the US were scored at 3, 4, 6, and 6, respectively (KPMG International, 2018).
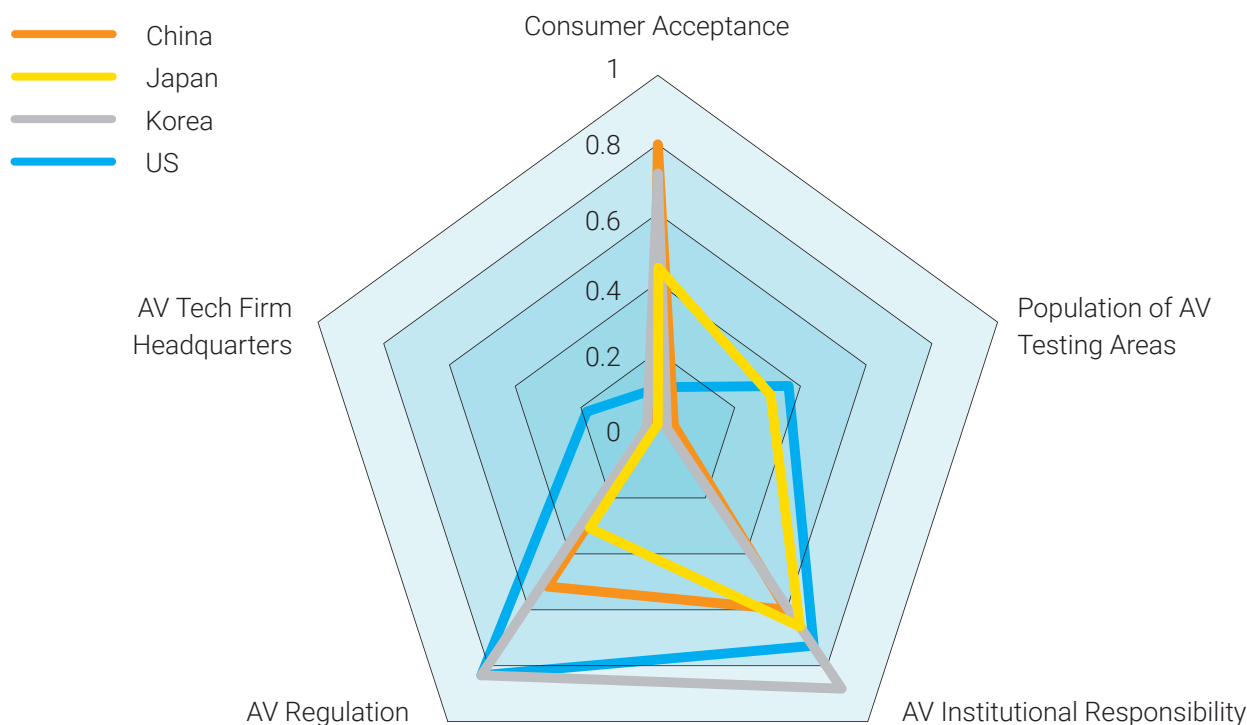
**Figure 4: Regulatory and social dimensions for autonomous vehicles**
*(Source: Made by the author based on the data from KPMG International (2019))*

to the index representing the number of AV technology firms' headquarters based on the KPMG International (2019), the US has the highest score of 0.176 followed by Korea (0.043). Japan is 0.029 while China (0.005) scored the lowest among the four countries (KPMG International, 2019).

In addition to AV regulations, social acceptance for AVs appears to be different among countries. As part of the consumer AI acceptance index, a consumer AV acceptance score—based on a branded research online consumer panel survey—shows that China scored the highest with 0.783 followed by South Korea's score of 0.725 (KPMG International, 2019). Japan and the US scored 0.442 and 0.103 respectively (KPMG International, 2019). In addition, the proportion of population living in AV testing areas (cities) vary because the numbers and areas of designated testing sites are different among countries. The US scored 0.355 for the highest percentage of people living in an AV testing area, followed by Japan with a score of 0.301; China and Korea scored 0.043 and 0.020 respectively (KPMG International, 2019).

The regulatory and social dimension scores of AV regulation for these four countries are compared in Figure 4.

The figure suggests that the US and Korea are very proactive and less restrictive about AVs, and have good institutional support for AV test driving. Japan is somewhat passive and cautious, with less institutional arrangement for AVs from the government. However, it is interesting to note that Korean consumers are the least receptive to AVs, and therefore test driving is limited to certain areas (smallest population living in test driving areas). Chinese and American consumers are highly receptive to AVs; particularly the US, as test driving is allowed in more areas than the three other countries, as indicated by the proportion of population in test areas. This suggests that the US is the least strict country when it comes to autonomous driving. It has not enacted specific legislation regarding AVs, but instead established guidelines based on SAE International standards that are used when establishing policies. In the US and Germany, AVs have already been put into operation on public roads.

Meanwhile, Japan has not yet passed legislation, but is preparing for Level-5 autonomous vehicle testing in advance of the Tokyo Olympics (Lee, 2018). Both China and Japan have declared their intentions to boost autonomous vehicle commercialization, and both have already passed related bills to allow test driving in limited areas. Additionally, Japan allows people to use cell phones while engaged in level 3 autonomous driving. Korea has also established a new law that addresses the commercialization of AVs, which is similar to the law for testing AVs. Despite the differences in regulating AVs, countries are similarly moving toward developing regulatory frameworks by introducing restrictions, limiting driving tests, and providing terms of technical standards. That said, there are still differences within these four countries' regulations in terms of technology-supported driving and safety measures.

## Conclusions and Policy Recommendations

As governments consider disruptive technologies as a source of future economic competitiveness, many have been shifting their regulatory positions from a regulatory paternalistic position to a somewhat deregulatory position, as seen in sandbox initiatives. While the regulation of disruptive technologies has weakened worldwide due to many people believing that regulation can harm the development of novel technologies, the risks and uncertainties associated with disruptive technologies still remain valid and require some form of regulation. At the same time, ethical guidelines often precede specific and formal

regulations due to the uncertain nature of those novel technologies. This study suggests there are two distinctive approaches—an ethical approach and legal/regulatory approach to new disruptive technologies. Examining the ethical guidelines of AI and the regulatory positions of AVs, this study suggests an ethical approach as an informal and unofficial guideline with key principles, which is often introduced before specific and formal regulations are adopted by governments. The ethical approach offers a broad range of key values to be considered for the design, development, deployment, and use of particular disruptive technologies. This study also suggests that regulatory decisions on disruptive technologies are often affected by uncertainties regarding the expected outcomes and social risk tolerance in relation to a specific technology. The regulatory positions of different countries might vary, primarily because of the expected roles of governments and market competition.

Regulatory schemes for novel technologies are not necessarily different from conventional technologies in a society, because regulatory politics are often similarly applied, regardless of the type of technology. However, we believe that disruptive technologies might create new regulatory dynamics in a country because of their novelties as well as their social risks and perceived uncertainty. Considering the implications of ethical and regulatory approaches, as well as their strengths and weaknesses, societies must manage disruptive technologies by carefully adopting and designing both approaches in order to address their uncertainties and perceived social risk. The following recommendations are proposed:

**Recommendation 1:** Moralizing disruptive technologies should precede, and should be fully discussed and shared among different stakeholder prior to regulating them. Before a society adopts and enacts specific regulatory frameworks for disruptive technologies, ethical guidelines (i.e., AI principles or AI ethical guidelines) must be jointly formulated based upon a thorough deliberation of particular disruptive technologies by different stakeholders representing industries, researchers, consumers, NGOs, international organizations, and policymakers.

**Recommendation 2:** AI ethical guidelines should support sustainable and human-centric societies by minimizing the negative socio-economic and international consequences of disruptive technologies (i.e., inequality, unemployment, psychological problems, etc.), while maximizing their potential benefits for environmental sustainability, quality of life among others.

**Recommendation 3:** Once a general consensus is made on general ethical guidelines, they should be elaborated and specified in details targeting individual stakeholder groups representing different actors and sectors. Specific AI ethical guidelines should be developed and customized for AI designers, developers, adopters, users, etc. based on the AI lifecycle. In addition, industry and sector specific ethical guidelines should be developed and applied to each sector (care industry, manufacturing industry, service industry, etc.).

**Recommendation 4:** In regulating AI and other disruptive technologies, governments should align regulations with key values and goals embedded in various AI ethical guidelines (transparency, trustworthiness, lawfulness, fairness, security, accountability, robustness, etc.) and aim to minimize the potential social risks and negative consequences of AI by preventing and restricting possible data abuses or misuses, ensuring fair and transparent algorithms, in addition to establishing institutional and financial mechanisms through which the negative consequences of AI are systematically corrected.

**Recommendation 5:** Governments should ensure the quality of AI ecosystems by increasing government and non-government investment in R&D and human resources for AI by maintaining fair market competition among AI-related private companies, and by promoting AI utilities for social and economic benefits.

**Recommendation 6:** Governments should carefully design and introduce regulatory sandbox approaches to prevent unnecessarily strict and obstructive regulations that may impede AI industries but also facilitate developing AI and exploring AI-related innovative business models.
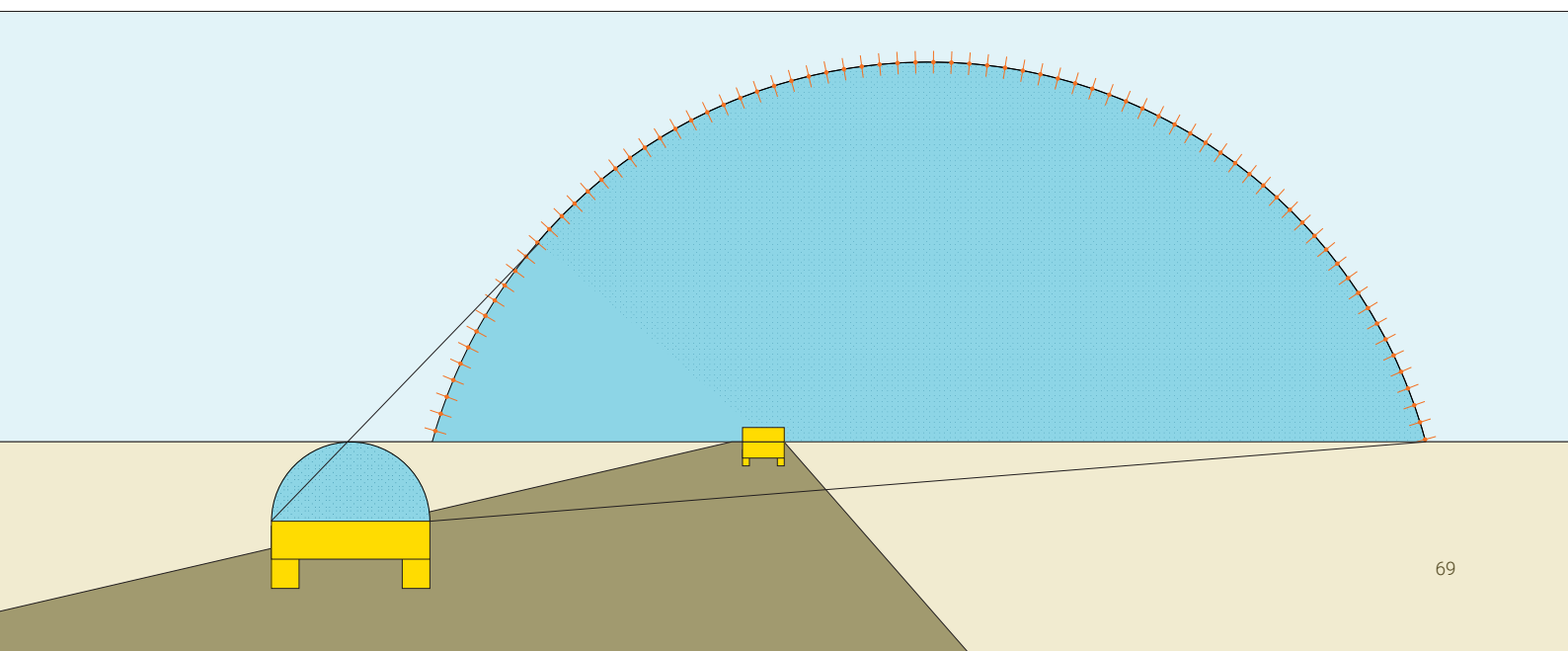
# References

Aghion, P., Algan, Y., Cahuc, P., & Shleifer, A. (2010). Regulation and distrust. *The Quarterly Journal of Economics,* 125(3), 1015-1049.

Chang, I. (2019). US Legislative Trends and Implications for Gene Editing Technology. *Study on The American Constitution,* 30(1), 213-242.

Choe, Y. S., & Jeong, J. (1993). Charitable Contributions by Low- and Middle-Income Taxpayers: Further Evidence with a New Method. *National Tax Journal,* 46, 33–39.

Beck, T., Levine, R., & Demirgüç-Kunt, A. (2002). *Law and finance: why does legal origin matter?* The World Bank.

Becker, G. S., & Stigler, G. J. (1974). Law enforcement, malfeasance, and compensation of enforcers. *The Journal of Legal Studies,* 3(1), 1-18.

Black, J. (1998). Regulation as Facilitation: Negotiating the Genetic Revolution. *Mod. L. Rev.,* 61, 621.

Berkhout, J., & Lowery, D. (2010). The changing demography of the EU interest system since 1990. *European Union Politics,* 11(3), 447-461.

Borges, B. J. P., Arantes, O. M. N., Fernandes, A., Broach, J. R., Fernandes, P., & Bueno, M. (2018). Genetically Modified Labeling Policies: Moving Forward or Backward? *Frontiers in bioengineering and biotechnology,* 6, 181.

Brodsky, J. S. (2016). Autonomous vehicle regulation: How an uncertain legal landscape may hit the brakes on self-driving cars. *Berkeley Tech., LJ,* 31, 851.

Brunel, C., & Levinson, A. (2013). Measuring Environmental Regulatory Stringency. *OECD Trade and Environment Working Papers,* 2013(5), 0_1.

Castor, A. (11 May 2018). *How Japan Is Creating a Template for Cryptocurrency Regulation. Bitcoin magazine.* https://bitcoinmagazine.com/articles/how-japan-creating-template-cryptocurrency-regulation

Charo, R. A. (2016). The legal and regulatory context for human gene editing. *Issues in Science and Technology,* 32(3), 39.

Charron, N., Dahlström, C., & Lapuente, V. (2012). No law without a state. *Journal of Comparative Economics,* 40(2), 176-193.

Cheon, C. (2018). *Global ICO Regulation Trends and Implications.* Korea Capital Market Institute Issue report, 18-06.

Cohen, J. (19 March 2019). WHO panel proposes new global registry for all CRISPR human experiments, *Science.* https://www.sciencemag.org/news/2019/03/who-panel-proposes-new-global-registry-all-crispr-human-experiments

ComplyAdvantage. (2018). Cryptocurrency Regulations Around The World. https://complyadvantage.com/blog/cryptocurrency-regulations-around-world

Cook, K., Shortell, S. M., Conrad, D. A., & Morrisey, M. A. (1983). A theory of organizational response to regulation: the case of hospitals. *Academy of Management Review,* 8(2), 193-205.

Cyranoski, D. (2016). *CRISPR gene-editing tested in a person for the first time.* Nature. doi:10.1038/nature.2016.20988

Cyranoski, D., & Ledford, H. (2018). *Genome-edited baby claim provokes international outcry.* Nature. https://www.nature.com/articles/d41586-018-07545-0

Cyranoski, D. (2019). *China to tighten rules on gene editing in humans.* Nature. https://www.nature.com/articles/d41586-019-00773-y

Cyranoski, D. (2019). *China announces hefty fines for unauthorized collection of DNA.* Nature. https://www.nature.com/articles/d41586-019-01868-2

Cyranoski, D. (2019). *Japan approves first human-animal embryo experiments.* Nature. https://www.nature.com/articles/d41586-019-02275-3

Das, S. (2017). *China's Central Bank Completes Digital Currency Trial on a Blockchain,* CCN. https://www.ccn.com/chinas-central-bank-completes-digital-currency-trial-blockchain

De Bruycker, I., & Beyers, J. (2015). Balanced or biased? Interest groups and legislative lobbying in the European news media. *Political Communication,* 32(3), 453-474.

Deng, C. (2018, January 11). *China Quietly Orders Closing of Bitcoin Mining Operations,* The Wall Street Journal. https://www.wsj.com/articles/china-quietly-orders-closing-of-bitcoin-mining-operations-1515594021

Dickinson, S. (2018, July 17). *Self Driving Cars in China: The Absence of Non-Technical Barriers,* China Law Blog. https://www.chinalawblog.com/2018/07/self-driving-cars-in-china-the-absence-of-non-technical-barriers.html

Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2002). The regulation of entry. *The quarterly Journal of economics,* 117(1), 1-37.

Downey, H. K., & Slocum, J. W. (1975). Uncertainty: Measures, research, and sources of variation. *Academy of Management journal,* 18(3), 562-578.

European Central Bank. (2012). *Virtual Currency Schemes.* https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemes201210en.pdf

European Parliament. (2018). Report on three-dimensional printing, a challenge in the fields of intellectual property rights and civil liability (2017/2007(INI)).

Fleming, L. (2001). Recombinant Uncertainty in Technological Search. *Management Science,* 47(1), 117-132.

Fordham, B., & McKeown, T. (2003). Selection and Influence: Interest Groups and Congressional Voting on Trade Policy. *International Organization,* 57(3), 519-549.

Gai, P., Kemp, M., Sánchez Serrano, A., & Schnabel, I. (2019). *Regulatory complexity and the quest for robust regulation* (No. 8). European Systemic Risk Board.

Glaeser, E. L., & Shleifer, A. (2002). Legal origins. *The Quarterly Journal of Economics,* 117(4), 1193-1229.

Globus, R., & Qimron, U. (2018). A technological and regulatory outlook on CRISPR crop editing. *Journal of cellular biochemistry,* 119(2), 1291-1298.

Go, J. (December 2). *[Genetic Editing Baby Controversy] a rekindled debate on human embryo research.* Dong-A Science. http://dongascience.donga.com/news.php?idx=25463

Hacker, P., & Thomale, C. (2018). Crypto-Securities Regulation: ICOs, Token Sales and Cryptocurrencies under EU Financial Law. *European Company and Financial Law Review,* 15(4), 645-696.

Hail, L., & Leuz, C. (2006). International differences in the cost of equity capital: Do legal institutions and securities regulation matter?. *Journal of accounting research,* 44(3), 485-531.

Hofstede, G. (2015). *The 6-D model of national culture.* https://geerthofstede.com/culture-geert-hofstede-gert-jan-hofstede/6d-model-of-national-culture/

Houben, R., & Snyers, A. (2018). Cryptocurrencies and blockchain: legal context and implications for financial crime, money laundering and tax evasion. Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament.

Hunt, G., & Mehta, M. (Eds.) (2013). *Nanotechnology: "Risk, Ethics and Law".* Routledge.

Hwang, Y. (2018, August 29). *Deregulation of human embryo research and other...There's going to be a controversy over bioethics.* http://www.hani.co.kr/arti/society/health/859834.html

Jeung, T. (2018, October 10). *Japan Is Drafting a Rulebook for Ethically Editing the Genes of Human Embryos: Which country will be first to create a CRISPR baby?* https://www.inverse.com/article/49725-governments-regulate-human-embryo-gene-editing

Kahan, D. M., Jenkins-Smith, H., & Braman, D (2011). Cultural cognition of scientific consensus. *Journal of risk research,* 14(2), 147-174.

Kerwer, D. (2005). Rules that many use: Standards and global regulation. *Governance, 18*(4), 611-632.

Kim, B. (2015). A Study on Uber Taxi and the Fit of It. *Chonbuk Law Review, 46, 99~*134.

Kisiel, D. (2018). Legal concept of internet currencies. *Financial Law Review,* 11(3), 81-91.

Kolacz, M. K., Quintavalla A., & Yalnazov. O. (2019). Who Should Regulate Disruptive Technology? European *Journal of Risk Regulation,* 10(1), 4-22.

KPMG International. (2018). *Autonomous Vehicles Readiness Index - Assessing countries' openness and preparedness for autonomous vehicles.* https://assets.kpmg/content/dam/kpmg/xx/pdf/2018/01/avri.pdf

KPMG International. (2019). *2019 Autonomous Vehicles Readiness Index – Assessing countries' preparedness for autonomous vehicles.* https://assets.kpmg/content/dam/kpmg/xx/pdf/2019/02/2019-autonomous-vehicles-readiness-index.pdf

Kun, L., & Xiaodong, W. (2019, March 1). *Rules to be revised on organ donations.* China Daily. http://www.chinadaily.com.cn/a/201903/01/WS5c78936aa3106c65c34ec237.html

Lander, E. S., Baylis, F., Zhang, F., Charpentier, E., Berg, P., Bourgain, C., Friedrich, B., Joung, J. K., Li, J., Liu, D., Naldini, L., Nie, J., Qiu, R., Schoene-Seifert, B., Shao, F., Terry, S., Wei, W., & Winnacker, E. (2019, March 19). *Adopt a moratorium on heritable genome editing,* Nature. https://www.sciencemag.org/news/2019/03/who-panel-proposes-new-global-registry-all-crispr-human-experiments

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Robert, V. (1997). Legal determinants of external finance. *Journal of Finance* 52, 1131–1150.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Robert, V. (1998). Law and finance. *Journal of Political Economy* 106, 1113–1155.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Robert, V. (1999). The quality of government. *Journal of Law, Economics, and Organization* 15, 222–279.

La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2008). The economic consequences of legal origins. *Journal of economic literature* 46(2), 285-332.

Lee, S. (2018). *Issues on Regulatory Reform for Industrial Revitalization of Self-driving Cars.* ICT Spot issue, IITP, S18-06.

Lee, S. & Kim, H. (2018). International Regulatory Trends on Genome Editing Research Using Human Embryo and Its Implication. *Korean Journal of Medicine and Law* 26(2), 71-96.

Marchant, G., Meyer, A., & Scanlon, M. (2010). Integrating social and ethical concerns into regulatory decision-making for emerging technologies. *Minn. JL Sci. & Tech.,* 11, 345.

Marris, C., Langford, I., Saunderson, T., & O'Riordan, T. (1997). Exploring the "psychometric paradigm": comparisons between aggregate and individual analyses. *Risk analysis, 17*(3), 303-312.

Marshall, A. (2018). *New York City Goes After Uber and Lyft.* Wired. https://www.wired.com/story/new-york-city-cap-uber-lyft

Martin-Laffon, J., Kuntz, M., & Ricroch, A. E. (2019). Worldwide CRISPR patent landscape shows strong geographical biases. *Nature biotechnology.* 37(6), 613-620.

Matsuda, D., Mears, E., & Shimada, Y. (2019). *Legalization of Self-Driving Vehicles in Japan: Progress Made, but Obstacles Remain.* DLA Piper. https://www.dlapiper.com/en/global/insights/publications/2019/06/legalization-of-self-driving-vehicles-in-japan

Milliken, F. J. (1987). Three types of perceived uncertainty about the environment: State, effect, and response uncertainty. *Academy of Management review,* 12(1), 133-143.

Ministry of Science and ICT. (2017). *Korea Provides Gene Scissors, U.S. Corrects Human Embryo Gene Mutation.* Press Releases.

Ministry of Science and ICT and KISTEP. (2018). Comparative analysis of domestic and foreign legislation on autonomous vehicles and policy alternatives, *In Science, ICT Policy and Technology Trends, 128.*

Molteni, M. (2019, July 30). The World Health Organization Says No More Gene-Edited Babies, *Wired*. https://www.wired.com/story/the-world-health-organization-says-no-more-gene-edited-babies

Nature. (2019, March 13) *Hybrid embryos, ketamine drug and dark photons.* https://www.nature.com/articles/d41586-019-00790-x

Normile, D. (2019). *China tightens its regulation of some human gene editing, labeling it 'high-risk'.* Science. https://www.sciencemag.org/news/2019/02/china-tightens-its-regulation-some-human-gene-editing-labeling-it-high-risk

Normile, D. (2019) *Gene-edited foods are safe, Japanese panel concludes,* Science. https://www.sciencemag.org/news/2019/03/gene-edited-foods-are-safe-japanese-panel-concludes

OECD. (2018). Blockchain Technology and Corporate Governance.

Ogus, A. (2005). Regulatory paternalism: when is it justified?. *Corporate governance in context: Corporations, states, and markets in Europe, Japan, and the US*, 303-320.

Ormond, K. E., Mortlock, D. P., Scholes, D. T., Bombard, Y., Brody, L. C., Faucett, W. A., Garrison, N. A., Hercher, L., Isasi, R., Middleton, A., Musunuru, K., Shriner, D., Virani, A., & Young, C. E. (2017). Human germline genome editing. *The American Journal of Human Genetics. 101*(2), 167-176.

Oshiro, Y., & Ohkohchi, N. (2017). Three-dimensional liver surgery simulation: computer-assisted surgical planning with three-dimensional simulation software and three-dimensional printing. *Tissue Engineering Part A, 23*(11-12), 474-480.

Park, T. (2019). *Does Uber want to tap the Korean market again? Foreign Taxi Call Service Initiated.* Hankyoreh. http://www.hani.co.kr/arti/economy/it/879525.html#csidx7f70c05e63c5236a29bda7b854ff47f

Pinto, C. (2012). How autonomous vehicle policy in California and Nevada addresses technological and non-technological liabilities. Intersect: *The Stanford Journal of Science, Technology, and Society,* 5.

Pollock, D. (2018, March 21). *G20 and Cryptocurrencies: Baby Steps Towards Regulatory Recommendations.* https://cointelegraph.com/news/g20-and-cryptocurrencies-baby-steps-towards-regulatory-recommendations

Herskind, N., Lim, C.K., & Hoist, S. (2019). How China will shape the future of autonomous vehicles. QVARTZ. https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles

Roca, J. B., Vaishnav, P., Morgan, M. G., Mendonça, J., & Fuchs, E. (2017). When risks cannot be seen: Regulating uncertainty in emerging technologies. *Research Policy, 46*(7), 1215-1233.

Sabel, C., Herrigel, G., & Kristensen, P. H. (2018). Regulation under uncertainty: The coevolution of industry and regulation. *Regulation & Governance, 12*(3), 371-394.

Schwinger, A. (2018, March 14). *Federal court holds that CFTC can regulate virtual currencies as commodities,* Norton Rose Fulbright website. https://www.nortonrosefulbright.com/en/knowledge/publications/6c7bcc30/federal-court-holds-that-cftc-can-regulate-virtual-currencies-as-commodities

Shim, M. (2019, June 13). Legal Issues Related to Genetics Patent. *Korea Institute of Intellectual Property.* https://www.kiip.re.kr/board/report/view.do?bd_gb=data&bd_cd=4&bd_item=0&po_item_gb=5&po_item_cd=&po_no=12504

Shukla-Jones, A., Friedrichs, S., & Winickoff, D. E. (2018). Gene editing in an international context: Scientific, economic and social issues across sectors. *OECD Science, Technology and Industry Working Papers,* 2018(4), 0_1-51.

Siegrist, M. (2010). Psychometric paradigm. *Encyclopedia of science and technology communication, Volume 2,* pp. 600-601. SAGE Publications.

Slovic, P. (1987). Perception of risk. *Science, 236*(4799), 280-285.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1982). Why study risk perception?. *Risk analysis, 2*(2), 83-93.

Starr, C. (1969). Social benefit versus technological risk. *Science,* 1232-1238.

Taeihagh, A., & Lim, H. S. M. (2019). Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Reviews, 39*(1), 103-128.

The Francis Crick Institute. (2019). Kathy Niakan: Human embryo genome editing licence. https://www.crick.ac.uk/research/labs/kathy-niakan/human-embryo-genome-editing-licence

The Law Library of Congress. (2014). Restrictions on Genetically Modified Organisms. Global Legal Research Center.

The Library of Congress. (2018, August 16). *Regulation of Cryptocurrency Around the World.* https://www.loc.gov/law/help/cryptocurrency/world-survey.php

The Library of Congress. (2018, August 16). *Regulation of Cryptocurrency: China.* https://www.loc.gov/law/help/cryptocurrency/china.php

The United Nations Economic and Social Council. (2017). Consolidated Resolution on the Construction of Vehicles (R E.3).

The US Congressional Research Service. (2019). H.R.3388 - SELF DRIVE Act. https://www.congress.gov/bill/115th-congress/house-bill/3388

Tomlinson, T. (2018). A crispr future for gene-editing regulation: a proposal for an updated biotechnology regulatory system in an era of human genomic editing. *Fordham L. Rev., 87*, 437.

Tzur, A. (2017). Uber Über regulation? Regulatory change following the emergence of new technologies in the taxi market. *Regulation & Governance.* https://doi.org/10.1111/rego.12170

UK Department for Transport. (2015). *The Pathway to Driverless Cars: A detailed review of regulations for automated vehicle technologies.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/401565/pathway-driverless-cars-main.pdf

Van Rijssen, W. J., & Morris, E. J. (2018). Safety and Risk Assessment of Food From Genetically Engineered Crops and Animals: The Challenges. *In Genetically Engineered Foods,* pp. 335-368. Academic Press.

Vienna Convention on Road Traffic. (2009). 1968 Vienna Convention on Road Traffic: Consolidated Resolution on Road Traffic. Revised on 14 August, 2009.

Wilson, J. (1980). *Politics of Regulation.* New York: Basic Books.

World Bank. (2016). *World Development Report 2016: Digital Dividends.* Washington, DC: World Bank. DOI:10.1596/978-1-4648-0671-1

World Economic Forum. (2017). *Global Competitiveness Index 2017-2018.* whttp://reports.weforum.org/global-competitiveness-index-2017-2018/?doing_wp_cron=1565516422.9761869907379150390625

Xinying, Z. (2019, March 1). Ministry to speed development of self-driving vehicles. http://www.chinadaily.com.cn/a/201903/01/WS5c78992ca3106c65c34ec27d.html
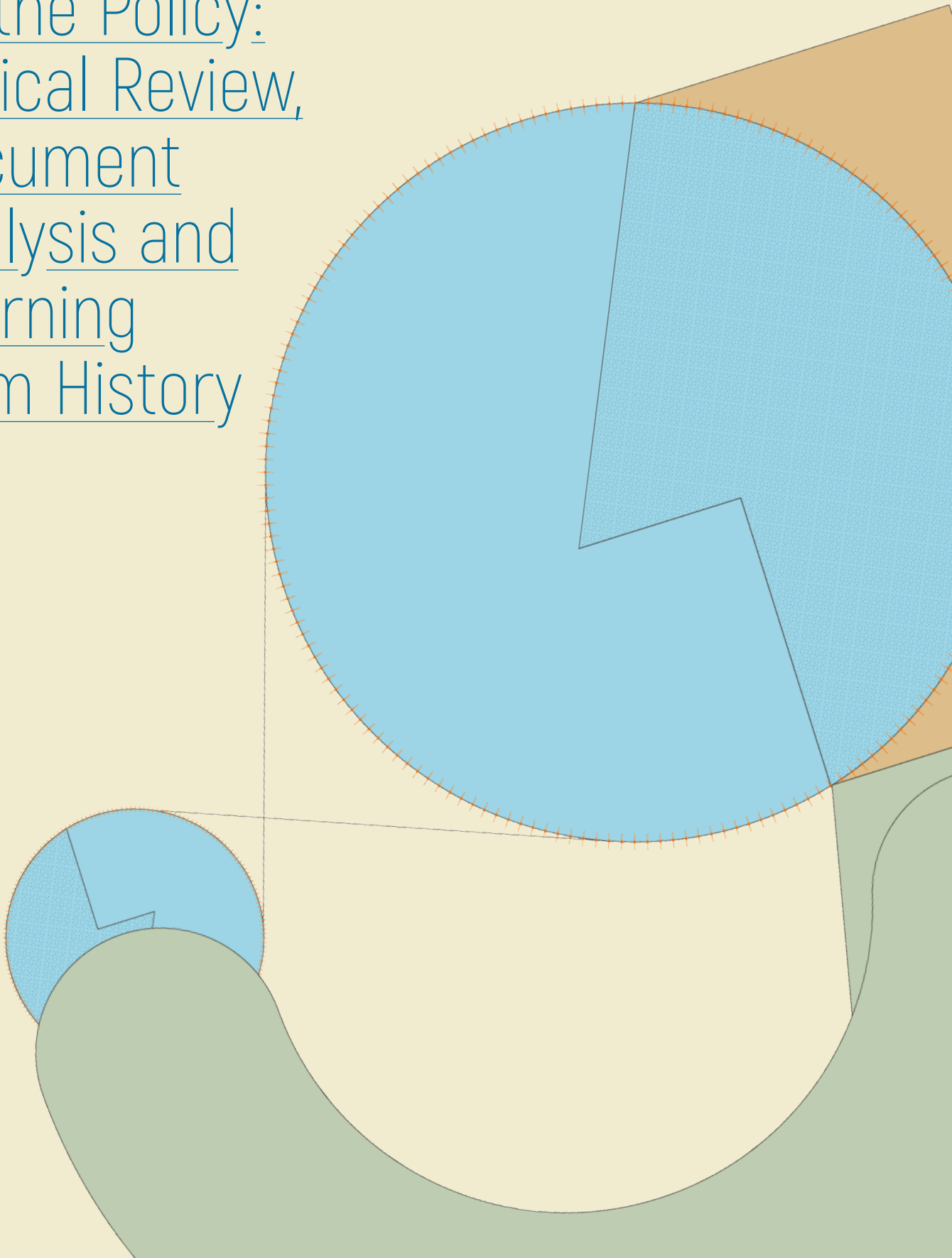
# Definition and Recognition of AI and its Influence on the Policy: Critical Review, Document Analysis and Learning from History

**Kyoung Jun Lee**
School of Management,
Kyung Hee University

**Yujeong Hwangbo**
Dept. of Social Network Science,
Kyung Hee University

## Abstract

Opacity of definitions hinders policy consensus; and while legal and policy measures require agreed definitions, to what artificial intelligence (AI) refers has not been made clear, especially in policy discussions. Incorrect or unscientific recognition of AI is still pervasive and misleads policymakers. Based on a critical review of AI definitions in research and business, this paper suggests a scientific definition of AI. AI is a discipline devoted to making entities (i.e., agents and principals) and infrastructures intelligent. That intelligence is the quality which enables entities and infrastructures to function (not think) appropriately (not humanlike) as an agent, principal, or infrastructure. We report that the Organization for Economic Co-operation and Development (OECD) changed its definition of AI in 2017 and how it has since improved from humanlike to rational and from thinking to action. We perform document analysis of numerous AI-related policy materials, especially dealing with the job impacts of AI, and find that many documents which view AI as a system that mimics humans are likely to overemphasize the job loss incurred by AI. Most job loss reports have either a "humanlike" definition, a "human-comparable" definition, or "no definition". We do not find "job loss" reports that rationally define AI, except for Russell (2019). Furthermore, by learning from history, we show that automation technology such as photography, automobiles, ATMs, and Internet intermediation did not reduce human jobs. Instead, we confirm that automation technologies, as well as AI, creates numerous jobs and industries, on which our future AI policies should focus. Similar to how machine learning systems learn from valid data, AI policy makers should learn from history to gain a scientific understanding of AI and an exact understanding of the effects of automation technologies. Ultimately, good AI policy comes from a good understanding of AI.

# 1. Scientific understanding of AI

How one recognizes something influences their attitude when dealing with it. With AI being a very new concept compared with traditional subjects such as physics, economics, and sociology, there have been numerous misunderstandings; and while these have been overcome by the AI communities themselves, there is still incorrect and unscientific recognition of AI. Definitional ambiguity hampers the possibility of conversation; and although legal and regulatory intervention requires agreed-upon definitions, consensus surrounding the definition of AI has been elusive, especially in policy conversations (Krafft et al., 2020). In the following sections, we attempt to correct this misconception, thereby redefining AI.

## 1.1. AI is a discipline not an entity

Although AI is a discipline, some view it as a physical thing, in other words, a machine or entity. For example, the physicist Stephen Hawking told the BBC that "[the] development of full artificial intelligence could spell the end of the human race" (Cellan-Jones, 2014). This statement highlights Stephen Hawking's misunderstanding of AI, which, in turn, can mislead mass media and people. Just as he regarded AI as an entity and not a discipline, the non-AI community and non-professional community sometimes show their misunderstanding of AI by defining it as "machines performing humanlike cognitive functions" (OECD, 2017) or "intellectual machines and systems… that could automatically sense people's situations or expectations, and offer necessary information before it is required" (Ema et al., 2016). That said, mainstream AI research communities have known AI is an activity devoted to making machines intelligent (Nilsson, 2010),[1] is the science of making machines smart (Hassabis, 2015), and is a discipline. The most frequently used textbook in AI, "Artificial Intelligence: A Modern Approach" (Russell & Norvig, 1995), says that AI is "one of the newest fields in science and engineering". Textbooks older than this also explain that AI is the study of how to make computers do things which, at the moment, people do better (Rich, Knight & Nair, 1985); the study of mental faculties through the use of computational models (Charniak & McDermott, 1985); and the study of the computations that make it possible to perceive, reason, and act (Winston, 1992).

## 1.2. AI is not about humans, it should be based on rationalism

The definition of AI should not include the word "human". Physics is not about humans, chemistry is not about humans; both are natural science. History is about humans, sociology is about humans; these are humanities and social science, respectively. AI is the science of the artificial (Simon, 1969), it is not a science about humans. A natural science similar to AI is brain science, which is concerned with how human and animal brains work. AI, however, is not about how the human brain works, since even animals can be intelligent. As such, AI should not deal solely with human intelligence. Including the word "human" in the definition of AI confines the scope of the discipline and misleads academic and practitioner communities. AI is simply an activity that makes certain entities intelligent. It is not about making machines humanlike in intelligence; Nor is it about making machines more intelligent than humans, despite numerous non-professionals explaining AI as trying to making something more intelligent than a human (Bostrom, 2014; Cellan-Jones, 2014; Clifford, 2017; Manyika et al., 2017; Niyazov, 2019; John, 2019; Adel, 2019).

We found evidence that even AI researchers such as Rich and Knight (1991), incorrectly define AI as about making humanlike intelligence or human-comparable intelligence. Defining AI as human-related is a very common mistake in the non-AI and non-professional communities, such as with the aforementioned OECD (2017) and Ema et al. (2016). Merriam-Webster also shows an incorrect understanding of AI by defining it as "the capability of a machine to imitate intelligent human behavior".

---

1. AI is the activity devoted to making machines intelligent, and intelligence is that quality which enables an entity to function appropriately and with foresight in its environment (Nilsson, 2010).

This misconception of AI as "imitating humans" comes from the misunderstanding of Alan Turing's imitation game, the so-called Turing Test. Alan Turing, the father of computer science, suggested using the test as an operational definition of a "machine that can think". If a machine can pass test, then he suggested we can say the machine can think. However, different from his original intention, early AI scholars considered passing the imitation game as the goal of AI. Many AI researchers began to think that the goal of AI was to make a machine that is indiscernible from a human.

However, this outdated belief began to change after Hayes and Ford's speech at the International Joint Conference on Artificial Intelligence (IJCAI) in Montreal, Canada in 1995. Hayes and Ford asserted that the Turing Test has harmed AI development. They explained how, to be able to fly, it is not necessary for us to construct a bird-like flying machine or a machine that is indiscernible from a bird. Just as aeronautics is based on Bernoulli equation (Bernoulli, 1738) and not ornithology, AI does not have to be based on brain science. Russell and Norvig (1995) also referred to Hayes and Ford (1995) in their famous book, "Artificial Intelligence: A Modern Approach".

They propose two dimensions on the view of AI: humanlike or rational and thinking or acting. In choosing rationality over humanlike and acting over thinking, theirs is the first really "modern" approach to AI in comparison with traditional textbooks. As will be discussed in the following sections, the AI community has evolved by overcoming the Turing Test and not emphasizing AI cognition. Gershman et al. (2015), also proposes computational rationality as a potential unifying paradigm for intelligence in brains, minds, and machines.

## 1.3. AI is not only about cognition

Certain explanations of AI emphasize the cognitive aspect (Drum, 2017; Miller-Merrell, 2019; Frey & Osborne, 2017; Manyika et al., 2017). For example, we see plenty of examples of using the word "cognitive" or "cognition" when defining AI, such as Eysenck et al.'s (1990) definition of AI as the "attempt to

develop complex computer programs that will be capable of performing difficult cognitive tasks". OECD (2017) also defines AI as "machines performing humanlike cognitive functions". Sometimes this emphasis on cognition stems from attempting to differentiate AI from robotics. However, robotics also deals with cognition. Bostrom's (2014) definition of superintelligence, as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest", also mistakenly emphasizes cognition. This emphasis on cognition is not only wrong but is also misleading, in that it implies the AI system can think. As Turing tried to explain, we cannot determine when a thing thinks or not. Instead, he simply suggested a proxy test for the decision. Emphasis on cognition runs the risk of neglecting the action aspect of AI, which is a more important aspect of intelligence.

The traditional explanation of intelligent systems says an intelligent system has three processes: perception, cognitive, and motor. The perceptual system consists of sensors and associated memories. The cognitive system receives information from the stores in its working memory and uses previously stored information in long-term memory to make decisions about how to respond. The motor system carries out the response (Card et al., 1983). However, this traditional sandwich (perception-cognitive-motor) model has been criticized, for example, by Hurley (1998), and has now evolved into "enactivism". This is defined as the manner in which a subject of perception creatively matches its actions to the requirements of its situation (Protevi, 2006). Similar to the relatively new enactivism, traditional behaviorism also excludes or doubts the central role of cognition in intelligent systems. As such, the view regarding cognition as the center of intelligence is now being challenged, such as in Auer-Welsbach (2019).[2] As explained above, there still exists a disagreement over the central role of cognition; hence, the definition of AI should not only include the word "cognitive".

---

2. The fundamental composition of the most advanced intelligent system, the Homo Sapiens system, is not comprised of independent information processing units which interface with each other via representations. Instead, the system is comprised of independent and parallel producers of activity which all interface directly with the world through perception and action, rather than interface with each other exclusively. From this perspective, the notions of central and peripheral systems evaporate, as everything is both central and peripheral.

## 1.4. AI should be extended to not just agents

To date, AI applications have been confined to making agents intelligent from the principal-agent perspective. Meaning that the agents in AI disciplines only refer to machines, software, and robots that are owned and controlled by human principals. For example, Nilsson's (2010) definition of AI, as explained earlier, satisfies all three conditions: (1) it is referred to as a discipline, (2) it is not humanlike, and (3) there is not only an emphasis on cognition. This definition is the most accepted and up-to-date, and is therefore referred to by the comprehensive review and prospect report, "Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence" (Stone et al., 2016).

However, Nilsson's (2010) definition has one limitation which confines the intelligent entity to only a machine. This similar to Hassabis' (2015) definition in its limitation. In this paper, we extend Nilsson's definition since AI now plays a wide role in society. It is important to remember that AI is a discipline which makes entities and infrastructures intelligent, whereby the entities not only refer to agents such as machines, but also include principals such as humans, organizations, businesses, and nations. Infrastructures include computing elements, which can be imbedded into the natural world such as forest, lakes, and seas, as well as artificial infrastructures such as roads, cities, buildings, and homes. The extension to infrastructures from entities in the definition of AI

removes the humanlike feature, since it is nonsense to imagine humanlike roads or buildings. We assume that the agent orientation in defining AI could lead to humanlike orientation, which we can avoid by extending the scope of AI in its definition.

At the time, Russell and Norvig's (1995) approach which defined AI as making rational agents was the most pioneering and scientific at the time, hence why their book has been the most widely used at top AI schools around the world for more than 20 years since its publication. That said, it is necessary to extend Nilsson's (2010) and Russel and Norvig's (1995) definition and approach from making agents rational to making entities and infrastructures rational. Until now, AI research has concentrated only on optimizing the behavior of agents under a given condition. However, sensors and their networking technologies, such as Internet of things (IoT) technology, and automatic recognition technologies, such as convolution neural networks (CNNs), enable making infrastructures intelligent. Nowadays, AI needs to deal with the intelligence of not only single entities but also of infrastructures. This enlarged perspective encompasses the efforts for and contributions to human intelligence augmentation. In other words, augmented intelligence and intelligence amplification (Licklider, 1960; Engelbart 1962).[3] Jordan (2018) suggests a new term called intelligent infrastructure (II). Our new AI definition encompasses intelligence amplification (IA) and II, as well as traditional agent-oriented AI.

---

3. By "augmenting human intellect" we mean increasing the capability of someone to approach a complex problem, to gain comprehension to suit their particular needs, and to derive solutions to the problem. In this respect, increased capability is taken to mean a mixture of the following: more rapid comprehension, better comprehension, the possibility of gaining a useful degree of comprehension in a situation that was previously too complex, speedier solutions, better solutions, and the possibility of finding solutions to problems that before seemed insoluble (Engelbart 1962).

# 2. Scientific definition of AI

The simplest definition of AI is *a discipline that makes entities and infrastructures intelligent.* If we refine that definition, AI is *a discipline devoted to making entities and infrastructures intelligent, with intelligence being that quality which enables entities and infrastructures to function appropriately.*

## 2.1. The meaning of functioning appropriately

"To function appropriately" is derived from Nilsson's (2010) definition. It also means "acting rationally", as per Russell and Norvig's (1995) two-by-two matrix. This paper will dispense with a detailed explanation of each quadrant of the matrix because we have already criticized humanlike and cognition emphases when defining AI in an earlier section. Appropriate functioning is necessary for an entity to survive and prosper. Intelligence is evolved for the process of survival and, simultaneously, becomes the result of the prospering of entities. Thus, appropriate functioning is developed through evolution for natural entities and through optimization by a designer for artificial agents and infrastructures. We found that Nilsson's (2010) "functioning appropriately" comes from Albus's (1991) definition of intelligence as "the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioral subgoals that support the system's ultimate goal". According to Albus (1991),

the criteria of success and the system's ultimate goal are defined externally to the intelligent system. For an intelligent machine system, the goals and success criteria are typically defined by designers, programmers, and operators. For intelligent biological creatures, the ultimate goal is gene propagation, with success criteria being defined by the processes of natural selection.

Albus (1991) deals with the intelligence of both artificial intelligent systems and intelligent nature. His notion of intelligence corresponds with Anastasi's (1992) explanation that intelligence is the combination of abilities required for survival and advancement within a particular culture, and with Roth and Dicke's (2005) definition of intelligence.[4] In the definition of AI, "appropriate action" is also found in Kubacki (2009).[5] The recognition of intelligence as an instrument for survival and prosperity has not been popular in AI communities, though the idea was prevalent in evolutionary biology and psychology. However, we can find attempts by AI communities who view AI for the survival and prosperity of entities. Weng (2002) regards the performance of an intelligent entity as keeping the norm defined by social groups,[6] which can be called "institutional intelligence". This approach can be called an institutional approach to AI. Since institutional economics is a relatively new discipline in economics, the institutional approach to AI is a novel area to investigate.

---

4. Intelligence may be defined and measured by the speed and success of how animals, including humans, solve problems to survive in their natural and social environments (Roth & Dicke 2005).

5. Artificial, "embodied" intelligence refers to the capability of an embodied "agent" to select an appropriate action based on the current, perceived situation (Kubacki 2009).

6. Different age groups of developmental robots have corresponding norms. If a developmental robot has reached the norm of a human group of age k, we can say that it has reached the equivalent human mental age k (Weng 2002).

## 2.2. Optimization as the science of functioning appropriately

AI traditionally focuses on optimizing the behaviors of an agent under the conditions and goals given by its principal. Intelligent agents fundamentally seek to form beliefs and plan actions in support of maximizing expected utility (Gershman et al., 2015). Our new definition of AI emphasizes approaches to enabling the appropriate actions of agents, principals, and infrastructures. Hence, AI can be divided into: (1) making agents rational – finding a method of optimizing the behavior of an agent with the goals given by the principal (i.e., the owner of the agent), and (2) making entities and infrastructures function appropriately – finding the optimization method in which the entities survive and prosper while interacting with other entities and the infrastructures in their environment by making the rational entities and infrastructures learn, adapt, and improve the institutions of the world or society. In either case, it is important to recognize that optimization is the main problem when creating such AIs.

Optimizing a behavior of an agent under a principal has been covered by many studies on optimization systems. It is important to note that there is an intractable problem in which the optimal solution cannot be obtained, no matter how good the computer's performance. Stuart Russell's recent book, "Human Compatible: Artificial Intelligence and Problem Control", also confirms that the existence of intractable problems gives us reason to think that computers cannot be as intelligent as humans. There is also no reason to assume that humans can solve intractable problems either (Russell, 2019).

Gershman et al. (2015) emphasizes that ideal maximizing expected utility (MEU) calculations may be intractable for real-world problems. That is, finding optimal solutions can be intractable, even though optimization can be effectively approximated by rational algorithms which maximize a more generally expected utility incorporating the costs of computation. Thus, even though AI methodology improves, there are still certain optimization problems which cannot be solved under limited time and resources.

Judd (1990) proved learning in neural networks is NP-complete, and thus demonstrated that it has no efficient general solution. Goodfellow et al. (2015) also confirmed neural networks cannot avoid local minima.[7] Google-developed quantum computers solved a problem in three minutes, while the IBM Summit, the most powerful supercomputer in existence, requires a calculation time of 10,000 years (Arute et al., 2019). If quantum computing, which is 1 billion times faster than current supercomputing, is well developed and widely used for optimizing problems, it may become possible to solve problems considered intractable. If so, the range of problems that mankind could solve would be drastically expanded. Russell (2019) confirms that quantum computation helps slightly in solving intractable problems, but not enough to change the basic conclusion that there is no reason to suppose that humans can solve intractable problems.

On the other hand, if such developments are not realized, AI will still be forced to incompletely solve numerous problems and create a system for making occasional mistakes. Such incomplete systems should be used safely under human control. Although the performance of deep learning algorithms has improved, mistakes (i.e., local optima) have not gone away, which is the main problem of deep learning. Since deep learning is simply a neural network, it inherits the characteristics of a neural network, such as inexplainability and error inevitability. Research into increasing explanatory possibilities continues, and automatic recognition by deep learning is evolving, however, there is still a danger due to recognition error. Therefore, it is only suitable for use in areas where mistakes are not fatal and statistically good results are achieved. Current AI methodology is essentially a system that is able to make mistakes (Szegedy et al., 2014; Nguyen et al., 2016). Thus, Facebook researchers (Bordes et al., 2015) emphasize research and development through artificial tasks, just as an artificial task, such as XOR (exclusive OR) (Minsky & Papert, 1969), led to the birth of a multi-layer perceptron (Rumelhart et al., 1986).

---

7. Do neural networks enter and escape a series of local minima? Do they move at varying speed as they approach and then pass a variety of saddle points? [...] we present evidence strongly suggesting that the answer to all of these questions is no (Goodfellow et al., 2015).

## 2.3. An AI approach defined as an optimization problem

An AI algorithm is an algorithm which can find an optimal path to a preferred goal node, provided that the heuristic function satisfies certain conditions (Hart et al., 1968). Genetic or evolutionary algorithms are a type of optimization algorithm, meaning they are used to find the maximum or minimum of a function (Carr, 2014) called a "fitness function" – often a black-box in real-world applications. Automated theorem proving also finds proofs via application of optimization methods (Yang et al., 2016).

Most machine learning problems, once formulated, can be solved as optimization problems, with the essence of most machine learning algorithms being to build an optimization model and learn the parameters in the objective function from the given data (Sun et al., 2019). Sun et al. (2019) formulates supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning as optimization problems. For example, with supervised learning, the goal is to find an optimal mapping function to minimize the loss function of the training samples. Deep learning, if without nonlinearity in the hidden layer, would reduce to a generalized linear model. As such, minimizing the nonlinear and nonconvex loss functions is difficult, and at best we seek good local optima (Efron and Hastie, 2016). Reinforcement learning is a branch of machine learning, whereby an agent interacts with the environment through a trial and error mechanism, and learns an optimal policy by maximizing cumulative rewards (Sutton and Barto, 1998). Dialogue can also be considered as optimal decision making (Gao et al., 2018). The goal of dialogue learning for realizing conversational AI is to find optimal policies to maximize expected rewards in a reinforcement learning framework.

## 2.4. Successful AI applications in the pursuit of optimization

Successful AI applications and developments include the optimization perspective in their explanations.

Libratus (Brown and Sandholm, 2017), the first AI system to defeat top humans in heads-up no-limit Texas hold 'em poker, formulates itself by finding the optimal strategy for solving subgames. While Libratus may not be able to arrive at an equilibrium by independently analyzing subtrees, it may be possible to improve the strategies in those subtrees when the original base strategy is suboptimal, as is typically the case when abstraction is applied. DeepMind's AlphaGo is also based on the optimization perspective, claiming that all games of perfect information have an optimal value function, which determines the outcome of the game from every board position or state, under perfect play by all players (David et al, 1986).

On the other hand, IBM's Watson is not based on the optimization perspective. Watson is a knowledge-based decision support tool that suffers from the requirement to manually craft and encode formal logical models of the target domain. This should be evolved into an interactive decision support capability that strikes a balance between a search system and a formal knowledge-based system (Ferrucci, 2012). IBM's Watson has not been successfully deployed, experiencing only failures, particularly in the medical field (Brown, 2017; Herper, 2017; Bloomberg, 2017; Strickland, 2019).

Softbank's Pepper is not formulated as an optimized machine either. As a result, Pepper is rather limited in how it can help customers and its answers do not seem that helpful (Mogg, 2018). Pepper's failure was predicted (Lee, 2014) and widely reported on (Alpeyev & Amano, 2016; Bivens, 2016; Boxall, 2017; Nichols, 2018). Hanson Robotics' robot, Sophia, is a typical example of AI being based on the incorrect humanlike perspective, rather than the rational optimization perspective. As such, it only makes jokes and cannot have meaningful conversations (Campanella, 2016). Similarly, Honda's ASIMO business operation has also been stopped (Ulanoff, 2018). Humanoids such as Pepper, Sophia, and ASIMO all failed because they were based on a humanlike paradigm and not on an optimization framework.

## 3. OECD's redefinition of AI

Of the aforementioned perspectives, the OECD (2017) definition of AI is the most inaccurate, as it includes all three misconceptions. OECD (2017) defined AI as "Machines performing humanlike cognitive functions", thereby mistaking AI as an entity and not a discipline and incorrectly believing that AI should be humanlike. When defining AI, OECD (2017) also only emphasized cognition – a common misconception. This critical mistake in the definition of AI by the world-leading policy organization could have resulted in misguided policy decisions. In 2017, OECD was advised by one of this paper's authors to revise its definition. Interestingly, OECD (2018) changed it to: "Equipping systems with cognitive functions that allow them to function appropriately and with foresight in their environment". From this, it is apparent that OECD (2018) adopted Nilsson's (2010) definition. In the new definition, OECD (2018) avoided the humanlike criterion, stating that AI is an activity, rather than simply objects such as machines. Unfortunately, OECD (2018) unnecessarily added the word "cognitive", meaning that even this definition was inaccurate. In 2019, the definition was revised again, removing the word "cognitive", to read: "An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy".

In the OECD (2019) definition, it is worth noting the phrase "given set of human-defined objectives". Since rationalization refers to optimization under human-defined objectives, the OECD (2019) definition can be seen as taking the "rational" perspective. It is also explained that AI technologies can only deliver value if they are part of the organization's strategy and are used in the right way (Hippold, 2019). This also corresponds to the phrase "given set of human-defined objectives" in OECD (2019). Gartner's criticism of AI misconceptions shows its "rational" approach to AI. It also criticizes humanlike AI, explaining that while some forms of AI might give the impression of being clever, it is unrealistic to think that current AI is similar or equivalent to human intelligence (Hippold, 2019).

## 4. Identifying the definition of AI's influence on policy: Document analysis

Through our document analysis we were able to find research that was very close to ours. Krafft et al. (2020) compares AI researchers' recognition of AI with policy reports' perspective of AI. Similar to our claim in this paper, Krafft et al. (2020) criticizes the human emphasis in the definition of AI in most AI policy reports, while noting that AI researchers' recognition is more inclined to rational emphasis. Krafft et al. (2020) found that 28% of definitions by AI researchers and 62% from published policy documents use the word "human". There was more disagreement over whether existential threats are relevant (42% agreed) – an issue more relevant to (hypothetical) humanlike AI. In our paper, we analyze AI policy-related reports and classified resources according to their definition or perspective on AI. We particularly focus on resources which define AI as humanlike (thinking or action) entities.

For the analysis, we had planned to perform document analysis to investigate their position on: (1) the concern, fear, peril, threat, and danger of AI; (2) the fairness of AI (discrimination, oppression, discrimination, and inequality); and (3) unemployment and job loss. However, it was difficult to obtain systematic results, since it is very time consuming to analyze the perspectives of reports only by human reading. At first, we considered automatic document analysis using AI techniques. However, it is still difficult to automate document analysis to replace human reading; although there is research on the subject, such as Hermann et al. (2015). In near future, AI-based document analysis software will help human researchers perform this kind of research. With such AI discipline-based software, human researchers will be able to improve their performance and reduce the necessary research time. During our research, because we could not find such software for our purposes, we narrowed our focus to only job related reports, then analyzed them by keyword search and human reading. Krafft et al.'s (2020) study also seems to be based on this method.

## 4.1. Relationship between the perception of AI and the expectation of job loss

We investigate the relationship between the perception of AI and the expectation of job loss incurred by AI. We conject that a policymaker who believes or defines AI as something that thinks or acts in humanlike manner will be likely to overemphasize AI's negative impact on job creation. We were able to find numerous reports using humanlike AI definitions, such as Miller-Merrell (2019), Molla (2019), and Hawksworth et al. (2018). For example, Miller-Merrell (2019) describes AI as a branch of computer science that uses machine learning algorithms which "mimic" cognitive functions, making machines more humanlike. While Molla (2019) explains machine learning as something that can make humanlike decisions.

## 4.2. AI-induced job loss expectation defining AI as humanlike

### Policy reports

The report "Australia's Future Workforce?" by the Committee for Economic Development of Australia (CEDA, 2015) recognizes the ability of computers to emulate human thought patterns, claiming that AI is able to take over intellectual tasks, as well as routine ones. Hindi (2017) argues that the real issue facing governments today is the failure to transition to a sustainable AI society, which will lead to massive job loss and economic downturn. Hindi (2017) defines AI as the ability for a machine to reproduce human behavior. Daniel (2020) asserts that the pace at which AI is replacing the way humans work, forecasts that the future to be fully automated, even to the extent that jobs for humans will no longer exist. She explains that intelligent AI-models are trained to enable them to "act like a human" in real-world situations and that machines "think like human minds".

### Business websites

Many business web sites also make similar mistakes. For example, John (2019) defines AI as computers or devices that mimic humanlike movements, and expects that with automation – the real essence of the AI revolution – robots will takeover of several jobs, although not all careers will be destroyed. Balatayan (2018) claims even white-collar jobs are being cut due to technological advancements, defining an AI system as any software that can mimic a rudimentary form of thinking.

McClelland (2020) explains that the impact of AI and automation will be profound, and that we need to prepare for a future where job loss reaches 99%. His definition of AI is based on the following two assumptions, that (1) we will continue making progress in building more intelligent machines, and (2) human intelligence arises from physical processes. With this in mind, McClelland (2020) concludes that we will build machines which have human-level or higher intelligence. However, these assumptions were criticized by George Zarkadakis in his seminal book, In Our Own Image. In it, he describes six metaphors that people have used over the past 2,000 years to try and explain human intelligence. Zarkadakis (2015) shows that each metaphor simply reflected the most advanced thinking of the time.

### Consulting and research institute reports

Bughin et al. (2017) at McKinsey define AI as the ability of machines to exhibit humanlike intelligence, and explains that AI-powered automation could have a profound impact on jobs and wages. The Digital Marketing Institute (2019) raises the question, of whether AI will really steal our jobs in the future, and characterizes AI systems as being able to do things that humans can do and imitate the way we think. Wisskirchen et al. (2017) of the IBA Global Employment Institute describes AI as the work processes of machines that would require intelligence if performed by humans, asserting that both blue-collar and white-collar sectors will be affected.

### Media reports

Dai and Jing (2018) of the South China Morning Post refers to Oxford-Yale AI impact research – based on a survey of 352 machine learning experts – which estimates that there is a 50% chance of AI outperforming humans in all tasks in just 45 years, and which could take over every job in the next

century. The research explains that AI is the science of "simulating" intelligent behavior in computers, enabling the latter to exhibit humanlike behavioral traits such as knowledge, reasoning, common sense, learning, and decision making. Knapton (2016) of the Telegraph reports that the rise of robots could lead to unemployment rates greater than 50%, and that many middle-class professionals' jobs would be outsourced to machines within the next few decades, leaving workers with more leisure time than ever. Such comments are common misconceptions of people who see AI as being humanlike. The report itself also uses the term humanlike robots. Kelly (2019) of Forbes maintains that AI, robotics, and technology will displace millions of workers, and defines AI as the ability of a machine to mimic human behavior.

Adel (2019) of Medium states that AI's effect on work will be disruptive, and predicts a future in which robots take jobs from human workers. Adel (2019) also defines AI as the act of "simulating the human brain" in a machine, i.e., creating an artificial human mind far more powerful than an actual human one. Wadhwa (2016) of FactorDaily argues that we are facing a jobless future because AI systems emulate the functioning of the human brain's neural networks. Xu (2017) of Northeastern's J-school's Ruggle Media reports that computers have become substitutes for various types of jobs for numerous reasons, such as recent developments in AI machine learning. Machine learning will not only reduce the huge demand for labor input with tasks since it can be routinized depending on pattern recognition, it will also increase the demand for labor-performing tasks that are not subject to computerization. Xu (2017) recognizes that every aspect of learning or any other feature of intelligence can, in principle, be so precisely described that a machine can be made to simulate it.

## 4.3. AI-induced job loss expectation regarding AI as a super-intelligent entity

Through the document analysis, we found a number of reports that regard AI as a competitor to humans, i.e., a superhuman entity. Although the reports do not explicitly describe AI as being humanlike, they also belong to the humanlike category. Cellan-Jones (2014) refers to Stephen Hawking's fears on the consequences of creating something that can match or surpass humans (who are limited by slow biological evolution), as well as the concerns that clever machines, capable of undertaking tasks performed by humans up until now, will swiftly destroy millions of jobs. Clifford (2017) refers to Elon Musk's belief that a machine could be far smarter than a human, that robots will be able to do jobs better than humans, and that there will certainly be job disruption. Manyika et al. (2017) of McKinsey is of a similar opinion, saying that "machines already exceed human performance". Finally, Niyazov (2019) assumes that AI algorithms and automated manufacturing are much better at performing tasks.

## 4.4. AI-induced job loss expectation without a specific definition of AI

There are also claims of job loss by AI without a specific definition of AI (Brynjolfsson & McAfee, 2011; Kurzweil Network, 2012; Frey & Osborne, 2013; World Economic Forum, 2016; Acemoglu & Restrepo, 2017; Frey & Osborne, 2017; Rieley, 2018; Lambert & Cone, 2019; Ambika, 2019; The Week, 2019; Muro et al., 2019). For example, Krafft et al. (2020) mentions that over 40% of policy reports do not have a definition of AI. Frey and Osborne (2013) of Oxford Martin School reports that 47% of total US employment is in the high-risk category, and that associated occupations are

potentially automatable over an unspecified number of years – perhaps a decade or two. The World Economic Forum (2016) holds that current trends could lead to a net employment impact of more than 5.1 million jobs lost to disruptive labor market changes from 2015–2020; with a total loss of 7.1 million jobs, two thirds of which are concentrated in the office and administrative job family, and a total gain of 2 million jobs in several smaller job families.

Using a model in which robots compete against human labor in various tasks, Acemoglu and Restrepo (2017) of the Massachusetts Institute of Technology (MIT) and Brown University show that robots may reduce employment and wages, and that the local labor market effects of robots can be estimated by regressing the change in employment and wages on the exposure to robots in each local labor market – defined from the national penetration of robots into each industry and the local distribution of employment across industries. Frey and Osborne (2017) of Oxford Martin School claim that recent developments in machine learning will put a substantial share of employment at risk across a wide range of occupations in the near future, and that nearly half of all US jobs were at risk from AI-powered automation. Rieley (2018) of the US Bureau of Labor Statistics also asserts that employment of bookkeepers is projected to decline 1.5% from 2016–2026, representing a loss of 25,200 jobs.

Ambika (2019) also maintains that AI technologies being adopted around the globe will replace numerous jobs currently being done by humans. The Week (2019) reports that over the next decade, automation and AI could put 54 million Americans out of work. Muro et al. (2019) of Brookings Institute reports that although robots are not replacing everyone, a quarter of US

jobs will be severely disrupted as AI accelerates the automation of existing work. Lambert and Cone (2019) of OxfordEconomics.com claim that with the rise of robots in business models, many sectors will be seriously disrupted and millions of existing jobs will be lost, with 20 million manufacturing jobs set to be lost to robots by 2030.

Most job loss reports have either a "humanlike" definition, a "human-comparable" definition, or "no definition". According to our definition of AI, we claim that job loss reports make mistakes due to the incorrect recognition and understanding of the characteristics of AI. We were unable to find job loss reports that define AI as rational, except for Russell (2019). Russell is a very respectable AI pioneer who wrote an innovative textbook on AI (Russell & Norvig, 1995). However, even though he makes an attempt, he confesses not to be qualified to opine on the job issue. Other AI experts, such as Lee (2018a), also make similar mistakes when defining AI by incorrectly emphasizing "humanlike" and "cognitive'". AI policies are too important to leave entirely to technical AI experts. As Russell (2019) asserts, the job issue is too important to leave entirely to economists. For example, Martin Ford, a journalist who is not an AI expert, wrote a book exaggerating job loss from AI (Ford, 2015). However, he seems to have changed his mind after interviewing numerous world-renowned AI experts (Ford, 2018). It is therefore necessary for us to explain AI to policy experts, as well as promote collaboration among AI and policy experts.

# 5. Automation creates more jobs than it eliminates: Learning from history

## 5.1. AI creates more jobs than it eliminates

Throughout the research, we found numerous reports claiming that AI will not eliminate jobs. Shrive (2018) claims that AI cannot replace humans in performing all tasks, especially in the property management domain. AI has been specifically developed to simplify repetitive and time-consuming processes, thereby freeing up time for property managers, letting agents, and contractors to deal with more pressing problems. Lokitz (2018) asserts that with every job taken over by a machine, there will be an equal number of opportunities for jobs to be done by people. Furthermore, in many cases, humans and machines will find themselves in symbiotic relationships, helping each other to do what they do best.

The World Economic Forum (2018) asserts that 38% of businesses surveyed expect to extend their workforce to new productivity-enhancing roles, more than a quarter expect automation to lead to the creation of new roles in their enterprise, about half of today's core jobs – making up the bulk of employment across industries – will remain stable up to 2022, and current estimates suggest a decline of 0.98 million jobs and a gain of 1.74 million jobs. Atkinson (2018) asserts that there is no reason to believe that this coming technology wave will be any different in pace and magnitude than previous waves. Each past wave has led to improved technology in a few key areas (e.g., steam engines, railroads, steel, electricity, chemical processing, and information technology), and these were then used by many sectors and processes. Within manufacturing, for example, each wave has led to important improvements, however, there have always been many other processes that have required human labor. The British Academy (2018) maintains

that while there is now a consensus that AI does not spell the end of work, neither will the transition be painless for all. Although human-level intelligence ('general AI') receives significant media attention, it is still some time away from being delivered, and it is unclear when it might be possible. Krafft et al. (2020) points out that hype surrounding general AI centers on humanlike AI, and that it is a problem that many policy analysts think of it in this way.

AdextAI (2019) explains that, as the technology has evolved, unemployment rates have decreased as a result of the new jobs created. Naudé (2019) holds that, in the foreseeable future, AI is unlikely to cause huge job losses (or job creation), at least in advanced economies. The main reasons for this conclusion are based on: (1) the fact that the methods used to calculate potential job losses are sensitive to assumptions; (2) automation may affect tasks more significantly, rather than the jobs within which they are performed; (3) net job creation can be positive because automation stimulates the creation of new jobs or jobs elsewhere; (4) diffusion of AI may be much slower than is thought or assumed; and (5) the tempo of innovation in AI is slowing down. Thomas (2019) explains that AI is poised to eliminate millions of current jobs and create millions of new ones – some of which have yet to be invented. Liang (2019) describes that recent advances in AI, while seemingly impressive, are very narrow in scope and require a lot of human supervision and input to work in real applications. While as many as 47% of current jobs contain tasks that may be automatable, less than 5% of jobs will be fully automatable by 2030. As with many new technologies that came before, AI tools will augment and not replace workers by automating subtasks of a job.

## 5.2. Automation proved more of a blessing than a threat

Garry Kasparov says that he is the first knowledge worker whose job was threatened by a machine (Knight, 2020). Referring to Kasparov, Knight (2020) claims that technology destroys jobs before creating new ones. This story has been repeated since the Industrial Revolution in the 19th century. For example, with the emergence and popularity of machines in 19th century Britain, many workers lost their jobs. Luddism centered around the defense of hand trades in the textile industry in the face of innovation which threatened jobs (Beckett, 2012). Led by artisans who felt their jobs were being threatened by the increased use of machines in the production process, Luddites began destroying machines as a form of protest. An agricultural manifestation of Luddism occurred during the Swing Riots of 1830, which saw the destruction of threshing machines. Although automation freed people from mundane and repetitive tasks, it caused some people to lose their jobs.

William Lee was an English clergyman and inventor who, in 1589, devised the first stocking frame knitting machine, the design of which was used for centuries. Having perfected his design and desiring to secure Queen Elizabeth I's patronage, whose partiality for knitted silk stockings was well known, Lee went to London to exhibit the loom before the Queen. However, her reaction was not what he had expected. She is said to have opposed the invention on the grounds that it would deprive a large number of poor people of their employment of hand knitting (Smiles, 2005).

Although people have always been afraid of new automation technologies, they always proved more of a blessing than a threat. As machine learning systems learn from data, intelligent human beings should learn from history. In 1790, 90% of Americans were farmers.

Nowadays that number is less than 2% (Dimitri et al., 2005). So, has American agriculture disappeared? The answer is no, it has simply become more automated. The US has transformed from an agricultural economy to an industrial economy, then to a service economy, and now to an information economy. Dimitri et al. (2005) concludes that automation creates far more jobs than it eliminates. Even if automation takes on a variety of professional roles, it does not always take away people's jobs.

## 5.3. The camera created more jobs and industries than it eliminated

Invented roughly 200 years ago, cameras began to be distributed about 100 years ago. At the time, many people thought that there would be no more need for artists as a result. However, cameras allowed for the development of modern art, and many painters used cameras in the studios. Even early contributors to the invention of photography and the camera were painters themselves, such as Leonardo da Vinci, who used the camera obscura for his painting, and Louis-Jacques-Mandé Daguerre, who was a theatre set painter and inventor of the daguerreotype process of photography (Daval, 1982). With cameras, i.e., the new automation technology of the time, painters were able to dramatically reduce the time needed for painting and sell photos of their works to more customers. The existing skills needed for drawing portraits, simply became the basis for becoming a better photographer. In other words, the new technology became an opportunity to expand the existing portrait market into the photography market (Benjamin, 1969).

In addition, the invention of camera allowed related industries to develop. New industries emerged, such as film manufacturing, camera manufacturing, film sales, photo album production, photo studios, photographic development, photo distribution,

newspapers, magazines, advertising, and publishing industries, etc. Cameras also contributed to the development of other industries. For example, as more people began to take cameras with them when they travelled, the photos being taken encouraged more people to travel. Cameras also had an impact on the movie industry (Jeong, 2015), while the influence of celebrities such as Marilyn Monroe and John F. Kennedy was greater as a result of photography. Today, not only do people take pictures with their smartphones, but the continued development of photography has created new businesses such as Facebook and Instagram.

## 5.4. Automobiles created jobs and industries

A photograph taken on 5th Avenue in New York in 1900 shows the horse and cart to be the predominant mode of transport. By 1913, in little more than a decade, the automobile had replaced the horse as the main form of transport. In turn, this led to the development of related industries, such as automobile manufacturers, mechanics, and automobile salesmen. In addition to the development of personal automobiles, the city bus, intercity bus, express bus, taxi, and trucking industries all developed. At the same time, the construction of roads and car parks resulted in an increase in jobs (Lee, 2018). Not only did automobiles spark a desire for long-distance travel, but by shortening travel times, the travel industry and related transportation, lodging, and restaurant industries also developed alongside one another.

## 5.5. Digital typesetting created more jobs by promoting publishing

Physical typesetting is the composition of text through the arranging of metal "types" and is most well-known in the production of newspapers in the late 19th century. Being a typesetter was a highly skilled position, so much so that when the Hankyoreh newspaper in Korea was founded in 1988, it was unable to find a skilled typesetter. To solve the problem, the newspaper introduced an innovative technology called the Computerized Typesetting System (CTS). Starting with the Hankyoreh newspaper, many newspapers in Korea soon adopted this system, leading to a lot of typesetters losing their jobs. At the same time, however, demand for digital typesetters increased, which the traditional typesetters quickly learned, becoming desktop publishing professionals (Lee et al., 2012).

## 5.6. ATMs created jobs by contributing to bank expansion

When Automated Teller Machines (ATMs) were first invented in the 1970s, there were serious concerns about the layoffs of tellers. In the 1980s, US banks introduced ATMs to improve work efficiency, with the number of employees per branch decreasing to one third as a result. Between 1995 and 2010, the number of ATMs in the US surged from 100,000 to 400,000. However, there was no massive unemployment, since the number of bank branches increased by more than 40%. Furthermore, by 2015, the number of bank employees had increased from 250,000 to 500,000. As the introduction of ATMs reduced the cost of creating new branches, banks were able to expand and hire more employees than in the past. In addition, with ATMs replacing simple deposit and withdrawal

services, banks were able to focus on developing profitable financial products such as loan counselling and insurance. As a result, bankers were freed up to perform more important tasks than ever before. Not only were new jobs created when ATMs took over performing simple and repetitive tasks, bankers were able to take charge of tasks requiring high-level capabilities (James, 2015; Deloitte, 2018).

## 5.7. Internet intermediaries created jobs by reintermediation

Baen and Guttery (1997) predicted that increased use of the Internet and information technology would have a dramatic and negative impact on the real estate industry in terms of both income and employment levels. They argued that buyers and sellers with access to information available via the Internet would have no need for traditional "infomediaries", and that several other players in real estate support positions would also be disintermediated by the Internet. The authors predicted job losses in sectors directly related to real estate, including sales agents and developers, as well as sectors involved in the support of real estate transactions, such as legal services and banking. Muhanna and Wolf (2002) revisited Baen and Guttery's (1997) examination of technology's effect on the real estate industry and found that, in general, their most ominous predictions of income and employment loss have not materialized. In the years since their 1997 article, according to the Bureau of Labor's statistics, the real estate industry, like most sectors in the US, has experienced steady growth. Specifically, more workers were employed as real estate agents, developers, and legal service providers.

It is often argued that as electronic markets lower the cost of market transactions, traditional roles for intermediaries will be eliminated, leading to "disintermediation". Bailey and Bakos (1997) discuss the findings of an exploratory study of intermediaries in electronic markets which suggests that markets do not necessarily become disintermediated as they become facilitated by information technology. Middle businesses, functions, or people need to move up the food chain to create new value or face being disintermediated. However, the "reintermediation" opportunities are greater than the disintermediation perils (Tapscott, 1997). Yoon (2015) also explains that attention should be paid to reintermediation, where the value of brokerage functions has been recently created. There will be an opportunity to create new value for middlemen connecting consumers and suppliers.

These aforementioned examples show that new technology does not threaten the existence of someone's job. Just as a painter adapted to the invention of the camera and found a new job in a related field, so will it be the same in the case of AI. People currently engaged in fields such as health care, architecture, and law, where AI is expected to be applied, will acquire AI-related skills and take on new jobs.

# 6. Summary and Conclusion

Incorrect or unscientific understanding of AI is still pervasive and misleads policymakers. While ambiguity in definition has hampered conversation, legal and regulatory intervention requires agreed-upon definitions. However, consensus over the definition of AI has been elusive thus far, especially in policy conversations (Krafft et al., 2020). In this study, we reviewed numerous definitions of AI, and based on our critical review, we suggest a scientific definition of AI. Namely, that AI is a discipline devoted to making entities and infrastructures intelligent, with the intelligence being that quality which enables agents, principals, and infrastructures to function appropriately. We have observed how, since 2017, OECD has continued to update its definition of AI; and have noted how OECD has improved its definition from humanlike to rational and from thinking to action.

We investigated numerous AI-related policy documents, particularly those dealing with the impact of AI on jobs, and found that those which view AI as a system that mimics humans are likely to overemphasize job loss incurred by AI as an automation technology. In addition, most job loss reports have either a "humanlike" definition, a "human-comparable" definition, or "no definition". We were unable to find job loss reports that defined AI as rational. Through our historical review, we showed that automation technology, such as photography, automobiles, ATMs, and the Internet as an automatic intermediation technology, did not reduce human jobs. Instead, they created numerous jobs and industries. AI will also create a wide range of jobs and industries, on which our future AI policies should instead focus. Similar to how machine learning systems learn from valid data, AI policy makers should learn from history to gain a scientific understanding of AI and an exact understanding of the effects of automation technologies. Ultimately, good AI policy comes from a good understanding of AI.

We suggest four policy recommendations as follows:

**Recommendation 1:** Policy experts should be well educated about what AI is and what is really going on in the AI researches and businesses. Especially, AI should be considered as a discipline making entities and infrastructures intelligent, and the intelligence is that quality that enables agents, principals, and infrastructure to function appropriately. AI should not be considered as human-like or super-human system. Past AI policies based on the old paradigm should be rewritten.

**Recommendation 2:** Government should make program for educating the administrative officials, policy experts in public-owned research institute, and lawmakers in the national assembly.

**Recommendation 3:** Just as machine learning systems learn from data, policymakers should also learn from history and data. The positive impacts of automation technology should be recognized by policy makers and the new AI policy should be established based on the new recognition.

**Recommendation 4:** Government and society should recognize the characteristics of AI, as an optimization system, to have more public benefit, faster business outcomes and less risks from AI adoption.

## Acknowledgements

# References

Acemoglu, D., & Restrepo, P. (2017). Robots and Jobs: Evidence from US Labor Markets. *NBER Working Paper No. 23285*.

Adel, K. (2019). The Future of Jobs in Artificial Intelligence Era. *medium.com*. Retrieved from **https://medium.com/analytics-vidhya/the-future-of-jobs-in-artificial-intelligence-era-93e34c33c25f**

Adext AI. (2019). "How Many Jobs Will Be Lost Because of Artificial Intelligence?" Is the Wrong Question. *Adext AI*. Retrieved from **https://blog.adext.com/jobs-lost-artificial-intelligence/**

Albus, J. (1991). Outline for a theory of intelligence. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3).

Alpeyev, P., & Amano, T. (2016). A Japanese Billionaire's Robot Dreams Are on Hold. *Bloomberg*. Retrieved from **https://www.bloomberg.com/news/articles/2016-10-27/a-japanese-billionaire-s-robot-dreams-are-on-hold**

Anastasi, A. (1992). What Counselors Should Know About the Use and Interpretation of Psychological Tests. *Journal of Counseling & Development, 70*(5).

Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., . . . Collins, R. (2019). Quantum supremacy using a programmable superconducting processor. *Nature, 574*(7779), 505-510.

Atkinson, R. (2018). Shaping structural change in an era of new technology. *policynetwork.org*.

Auer-Welsbach, C. (2019). Interview with cognitive scientist Newton Howard on AI. *Medium*.

Baen, J. S., & Guttery, R. S. (1997). The Coming Downsizing of Real Estate: Implications of Technology. *The Journal of Real Estate Portfolio Management, 3*(1), 1-18.

Bailey, J. P., & Bakos, Y. (1997). An exploratory study of the emerging role of electronic intermediaries. I*nternational Journal of Electronic Commerce, 1*(3), 7-20.

Baltayan, A. (2018). Robots, Automation & Technology Taking Over – Is Your Job at Risk? *Money Crashers*. Retrieved from **https://www.moneycrashers.com/robots-automation-technology-replacing-jobs/**

Beckett, J. (2012). *Luddites*. Retrieved from The Nottinghamshire Heritage Gateway: **http://www.nottsheritagegateway.org.uk/people/luddites.htm**

Bellman, R. E. (1978). *An Introduction to Artificial Intelligence: Can Computers Think?* San Francisco: Boyd & Fraser Pub. Co.

Benjamin, W. (1969). *Das kunstwerk im zeitalter seiner technischen reproduzierbarkeit.* Frankfurt am Main: Suhrkamp.

Bernoulli, D. (1738). *Hydrodynamica.* France: The University of Strasbourg: Johann Reinhold Dulsecker.

Bessen, J. (2015). Toil and Technology. *Finance and Development, 52*(1).

Biven, M. (2016). Pepper Salé: Lessons from the bitter Aldebaran / SoftBank project. Retrieved from **https://markbivens.com/m/archives/pepper-sale-lessons-from-the-bitter-aldebaran-softbank-project**

Bloomberg, J. (2017). **https://www.forbes.com/sites/jasonbloomberg/2017/07/02/is-ibm-watson-a-joke/.** *Forbes.*

Bordes, A., Weston, J., Chopra, S., Mikolov, T., Joulin, A., Rush, S., & Bottou, L. (2015). Artificial Tasks for Artificial Intelligence. *Facebook AI Research ICLR.*

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* UK: Oxford University Press.

Boxall, A. (2017). Pepper is everywhere in Japan, and nobody cares. Should we feel bad for robots?

British Academy. (2018). The impact of artificial intelligence on work. *royalsociety.org.*

Brown, J. (2017). Why Everyone Is Hating on IBM Watson—Including the People Who Helped Make It. *GIZMODO.* Retrieved from **https://gizmodo.com/why-everyone-is-hating-on-watson-including-the-people-w-1797510888**

Brown, N., & Sandholm, T. (2017). Safe and Nested Endgame Solving for Imperfect-Information Games. *In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence.*

Brynjolfsson, E., & McAfee, A. (2011). *Race against the machine: how the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy.* Digital Frontier Press.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., . . . Trench, M. (2017). Artificial intelligence: the next digital frontier? *McKinsey and Company Global Institute,* 1-80.

Choudhury, A. (2019). AI May Kill These 5 Jobs By 2030, Say Experts. *Analytics India Magazine.* Retrieved from **https://analyticsindiamag.com/ai-may-kill-these-5-jobs-by-2030-say-experts/**

Campanella, E. (2016). *Meet Sophia, the human-like robot that wants to be your friend and 'destroy humans'.* Retrieved from Global News: https://globalnews.ca/news/2888337/meet-sophia-the- human-like-robot-that-wants-to-be-your-friend-and-destroy-humans/

Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human-Computer Interaction.* Hillsdale.

Carr, J. A. (2014). An Introduction to Genetic Algorithms. *Senior Project, 1*(40), 7.

CEDA. (2015). Australia's future workforce? *CEDA.*

Cellan-Jones, R. (2014). *Stephen Hawking warns artificial intelligence could end mankind.* Retrieved from BBC: https://www.bbc.com/news/technology-30290540

Charniak, E., & McDermott, D. (1985). *Introduction to Artificial Intelligence.* Unitted States: Addison-Wesley Longman Publishing Co., Inc.

Clifford, C. (2017). *Elon Musk: 'Robots will be able to do everything better than us'.*

Dai, S., & Jing, M. (2018). *Worried AI will replace your job?Here's an explainer to prepare for that day.* Retrieved from SCMP: https://www.scmp.com/tech/innovation/article/2131339/worried-ai-will-replace-your-jobheres-explainer-prepare-day#:~:text=Tech%20%2F%20Innovation-,Worried%20AI%20will%20replace%20your%20job%3FHere's%20an,to%20prepare%20for%20that%20day&text=The%20Oxford%2D

Daniel, E. (2020). *Role of Artificial Intelligence in Human Revolution.* Retrieved from Thrive Global: https://thriveglobal.com/stories/role-of-artificial-intelligence-in-human-revolution/

Daval, J.-L. (1982). *Photography History of an Art.* First American Edition.

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). MIT Press.

Deloitte. (2018). *The Future of Work.* One and One Books.

Digital Marketing Institute. (2019). *The Rise of AI: Will It Take or Create Digital Jobs?* Retrieved from DMI Blog: https://digitalmarketinginstitute.com/blog/the-rise-of-ai-will-it-take-or-create-digital-jobs

Dimitri, C., Effland, A., & Conklin, N. (2005). The 20th Century Transformation of U.S. Agriculture and Farm Policy. *Economic Information Bulletin,* 17.

Drum, K. (2017). *You Will Lose Your Job to a Robot—and Sooner Than You Think.* Retrieved from Mother Jones: https://www.motherjones.com/politics/2017/10/you-will-lose-your-job-to-a-robot-and-sooner-than-you-think/

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science.* UK: Cambridge University.

Ema, A., Akiya, N., Osawa, H., Hattori, H., Oie, S., Ichise, R., & Kanzaki, N. (2016). Future Relations between Humans and Artificial Intelligence: A Stakeholder Opinion Survey in Japan. *IEEE Technology and Society Magazine, 35*(4), 68-75.

Engelbart, D. C. (1962). *Augmenting Human Intellect: A Conceptual Framework.* Stanford Research Institute.

Eysenck, M. W., Hunt, E., Ellis, A., & Johnson-Laird, P. N. (1991). T*he Blackwell Dictionary of Cognitive Psychology.* UK: Wiley-Blackwell.

Ferrucci, D. (2012). ntroduction to "This is Watson". *IBM Journal of Research and Development.*

Ford, M. (2015). *Rise of the Robots: Technology and the Threat of a Jobless Future.* Basic Books.

Ford, M. (2018). *Architects of Intelligence: The Truth about AI from the People Building it.* Packt Publishing Ltd.

Frey, C. B., & Osborne, M. A. (2013). The Future of Employment: How Susceptible are Jobs to Computerisation? *The Oxford Martin Programme on Technology and Employment.*

Frey, C. B., & Osborne, M. A. (2017). The Future of Employment: How Susceptible are Jobs to Computerisation? *Technological Forecasting and Social Change, 114,* 254-280.

Gao, J., Galley, M., & Li, L. (2018). *Neural Approaches to Conversational AI.* Retrieved from https://arxiv.org/pdf/1809.08267.pdf

Gershman, S. J., Horvitz2, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273-278.

Goodfellow, I. J., Vinyals, O., & Saxe, A. M. (2015). Qualitatively characterizing neural network optimization problems. *arXiv.*

Haugeland, J. (1985). *Artificial intelligence: The very idea.* MIT Press.

Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics, 4*(2), 100-107.

Hassabis, D. (2015). *DeepMind Technologies - The Theory of Everything.* Retrieved from Google Zeitgeist.

Hawksworth, J., Berriman, R., & Cameron, E. (2018). *Will robots really steal our jobs? An international analysis of the potential long term impact of automation.* PwC.

Hayes, P., & Ford, K. M. (1995). *Turing Test Considered Harmful.* International Joint Conference on Artificial Intelligence.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2014). Teaching Machines to Read and Comprehend. *In Advances in neural information processing systems,* 1693-1701.

Herper, M. (2017). *MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine.* Retrieved from Forbes: https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/#395dbe613774

Hindi, R. (2017). *How my research in AI put my dad out of a job.* Retrieved from Medium: https://medium.com/snips-ai/how-my-research-in-ai-put-my-dad-out-of-a-job-1a4c80ede1b0

Hippold, S. (2019). *Gartner Debunks Five Artificial Intelligence Misconceptions.* Retrieved from Gartner: https://www.gartner.com/en/newsroom/press-releases/2019-02-14-gartner-debunks-five-artificial-intelligence-misconce

Hurley, S. (1998). *Consciousness in Action.* United States: Harvard University Press.

Jeong, M. S. (2015). *Humanities travel through film.* Kyungsung University.
*12 Jobs that Will Be Soon Replaced by AI.* (2019). Retrieved from apruve: https://blog.apruve.com/12-jobs-that-will-be-soon-replaced-by-ai

Jordan, M. (2019). Artificial Intelligence—The Revolution Hasn't Happened Yet. *Harvard Data Science Review.*

Judd, J. S. (1990). *Neural Network Design and the Complexity of Learning.* MIT Press.

Kelly, J. (2019). *Unbridled Adoption of Artificial Intelligence May Result in Millions of Job Losses and Require Massive Retraining for Those Impacted.* Retrieved from Forbes: https://www.forbes.com/sites/jackkelly/2019/09/30/unbridled-adoption-of-artificial-intelligence-may-result-in-millions-of-job-losses-and-require-massive-retraining-for-those-impacted/#6cd5dac51de7

Knapton, S. (2016). *Robots will take over most jobs within 30 years, experts warn.* Retrieved from The Telegraph: https://www.telegraph.co.uk/news/science/science-news/12155808/ Robots-will-take-over- most-jobs-within-30-years-experts-warn.html

Knight, W. (2020). *Defeated Chess Champ Garry Kasparov Has Made Peace with AI.* Retrieved from Wired: https://www.wired.com/story/defeated-chess-champ-garry-kasparov-made-peace-ai/

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2019). Defining AI in Policy versus Practice. *arXiv*.

Kubacki, J. (2009). Artificial intelligence. *SpringerLink.* Retrieved from SpringerLink.

Kurzweil, R. (1990). *The Age of Intelligent Machines.* MIT Press.

Kurzweil Network. (2012). *2 Billion Jobs to Disappear by 2030.* Retrieved from Kurzweil Accelerating Intelligence: https://www.kurzweilai.net/2-billion-jobs-to-disappear-by-2030#!prettyPhoto

Lambert, J., & Cone, E. (2019). How Robots Change the World. *Oxford Economics.*

Lee, K. (2014). Human Robot Era is Far yet. *Money Today.*

Lee, K.-F. (2018a). *AI Superpowers: China, Silicon Valley, and the New World Order.* Houghton Mifflin Harcourt.

Lee, Y. J., Jang, S. L., & Kim, W. J. (2012). *Gutenberg's Return .* Idambooks (Korean).

Lee, S. (2018b). *The Future of the 4th Industrial Revolution.* One and One Books (Korean).

Liang, J., Ramanauskas, B., & Kurenkov, A. (2019). *Job Loss Due To AI — How Bad Is It Going To Be?* Retrieved from Skynet Today: https://www.skynettoday.com/editorials/ai-automation-job-loss

Licklider, J. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics.*

Lokitz, J. (2018). *The future of work: How humans and machines are evolving to work together.* Retrieved from businessmodelsinc.com: https://www.businessmodelsinc.com/machines/

Luger, G. F., & Stubblefield, W. A. (1993). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving (2nd ed.).* Benjamin-Cummings Publishing Co., Inc.

Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., . . . Sanghvi, S. (2017). Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages. *McKinsey Global Institute,* 1-160.

McClelland, C. (2020). *The Impact of Artificial Intelligence - Widespread Job Losses.* Retrieved from iotforall.com: **https://www.iotforall.com/impact-of-artificial-intelligence-job-losses/**

Minsky, M., & Papert, S. (1969). *Perceptrons: an introduction to computational geometry.* MIT Press.

Miller-Merrell, J. (2019). *Resources, How Artificial Intelligence (AI) is Changing Human.* Retrieved from Randstad RiseSmart: https://www.randstadrisesmart.com/blog/how-artificial-intelligence-ai-changing-human-resources

Mogg, T. (2018). *Pepper the robot fired from grocery store for not being up to the job.* Retrieved from Digital Trends: **https://www.digitaltrends.com/cool-tech/pepper-robot-grocery-store/**

Molla, R. (2019). *"Knowledge workers" could be the most impacted by future automation.* Retrieved from Vox: **https://www.vox.com/recode/2019/11/20/20964487/white-collar-automation-risk-stanford- brookings**

Muro, M., Maxim, R., & Whiton, J. (2019). *Automation and Artificial Intelligence: How machines are affecting people and places.* Retrieved from Brookings Metropolitan Policy Program: **https://www.brookings.edu/research/automation-and-artificial-intelligence-how-machines-affect-people-and-places/**

Naudé, W. (2019). The Race against the Robots and the Fallacy of the Giant Cheesecake: Immediate and Imagined Impacts of Artificial Intelligence. I*ZA Discussion Paper no. 12218.*

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *arXiv*.

Muhanna, W. A. (2002). The Impact of E-Commerce on the Real Estate Industry: Baen and Guttery Revisited. *Journal of Real Estate Portfolio Management*(2), 141-152.

Nichols, G. (2018). *Robot fired from grocery store for utter incompetence.* Retrieved from ZDNet: **https://www.zdnet.com/article/robot-fired-from-grocery-store-for-utter-incompetence/**

Nilsson, N. J. (2010). *The Quest for Artificial Intelligence: A History of Ideas and Achievements.* UK: Cambridge University Press.

Niyazov, S. (2019). *How the Replacement of Blue-Collar Jobs by AI Will Impact the Economy.* Retrieved from iotforall.com: https://www.iotforall.com/how-ai-replacing-blue-collar-jobs-impact-economy/#:~:text=AI%20Is%20Emerging&text=The%20research%20conducted%20by%20Accenture,%25%20to%204.6%25%20by%202035.&text=This%20will%20boost%20exports%2C%20encourage,the%20US%20a%20manufac

OECD. (2017). OECD Science, Technology and Industry Scoreboard 2017. *OECD.*

OECD. (2018). AI: Intelligent machines, smart policies. *OECD Digital Economy Papers,* 0-33.

Protevi, J. (2006). *A Dictionary of Continental Philosophy.* United States: Yale University Press.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence.* Retrieved from OECD: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Rich, E., Knight, K., & Nair, S. (1985). *Artificial intelligence.* New York: McGraw-Hill.

Rich, E., Knight, K., & Nair, S. (1991). *Artificial Intelligence.* New York: McGraw-Hill.

Rieley, M. (2018). In the money: occupational projections for the financial industry. *Beyond the Numbers, 7*(16).

Roth, G., & Dicke, U. (2005). Evolution of the brain and intelligence. *TRENDS in Cognitive Sciences, 9*(5), 250-257.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning Internal Representations by Error Propagation.* MIT Press.

Norvig, P., & Russell, S. J. (1995). *Artificial Intelligence: A Modern Approach.* United States: Prentice Hall.

Russell, S. J. (2019). *Human Compatible: Artificial Intelligence and Problem Control.* Viking.

Schalkoff, R. J. (1990). *Artificial Intelligence Engine.* United States: McGraw-Hill, Inc.

Page, S. E. (2018). *The Model Thinker: What You Need to Know to Make Data Work for You.* United States: Basic Books.

Shrive, T. (2018). *AI will never replace jobs in the property market.* Retrieved from Finance Digest: https://www.financedigest.com/ai-will-never-replace-jobs-in-the-property-market.html#:~:text=AI%20mimics%20human%20behaviour%20and,%2C%20in%20theory%2C%20be%20automated.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. v., . . . Sutsk. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587).

Simon, H. A. (1969). T*he Sciences of the Artificial.* MIT Press.

Smiles, S. (2005). *Rev. William Lee, inventor of the Stocking Frame.* Retrieved from victorianweb.org: **http://www.victorianweb.org/technology/inventors/lee.html**

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., . . . Tambe, M. (2016). *Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence.* Report of the 2014 study panel, Stanford University.

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum, 56*(4), 24-31.

Sun, S., Cao, Z., Zhu, H., & Zhao, J. (2019). A Survey of Optimization Methods from a Machine Learning Perspective. *arXiv*.

Barto, A., & Sutton, R. S. (1998). *Reinforcement Learning: An Introduction.* MIT Press.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *arXiv*.

Tapscott, D. (1997). Strategy in the new economy. *Strategy & leadership, 25*(6), 8-15.

The Week. (2019). *Will you lose your job to a robot?* Retrieved from The Week: **https://theweek.com/articles/866339/lose-job-robot**

Thomas, M. (2019). *Artificial Intelligence's Impact on the Future of Jobs.* Retrieved from builtin.com: https://builtin.com/artificial-intelligence/ai-replacing-jobs-creating-jobs

Turing, A. (1950). *Computing Machinery and Intelligence.* Mind.

Ulanoff, L. (2018). *Say Hello to Our Disappointing Robot Future.* Retrieved from Medium: **https://medium.com/@LanceUlanoff/say-hello-to-our-disappointing-robot-future-e6f7d1d42e24**

Wadhwa, V. (2016). *We are heading towards a jobless future: is it good or bad?* Retrieved from Factor Daily: **https://factordaily.com/artificial-intellgence-automation-india-jobless-future/**

Weng, J. (2002). A theory for mentally developing robots. *Proceedings 2nd International Conference on Development and Learning,* 131-140.

Winston, P. H. (1992). *Artificial Intelligence (Third edition).* United States: Addison-Wesley Longman Publishing.

Wisskirchen, G., Biacabe, B. T., Bormann, U., Muntz, A., Niehaus, G., Soler, G., & Brauchitsch, B. v. (2017). *Artificial Intelligence and Robotics and Their Impact on the Workplace.* IBA Global Employment Institute.

World Economic Forum. (2016). *The Future of Jobs Report 2016.* World Economic Forum.

World Economic Forum. (2018). *The Future of Jobs Report 2018.* World Economic Forum.

Xu, J. (2017). Will human beings lose jobs due to AI? *Ruggles Media.*

Yang, L.-A., Liu, J.-P., Chen, C.-H., & Chen, Y.-p. (2016). Automatically Proving Mathematical Theorems with Evolutionary Algorithms and Proof Assistants. *arXiv.*

Yoon, S. (2015). *Digital Economy Leadership by Don Tapscott.* Retrieved from https://www.mk.co.kr/news/business/view/2015/11/1081476/

Zarkadakis, G. (2016). *In Our Own Image: Savior or Destroyer? The History and Future of Artificial Intelligence.* Pegasus Books.

# Institutional and Technological Design Development Through Use of Case Based Discussion

# Regulatory Interventions For Emerging Economies Governing The Use Of Artificial Intelligence In Public Functions

Arindrajit Basu,
Elonnai Hickok and
Amber Sinha
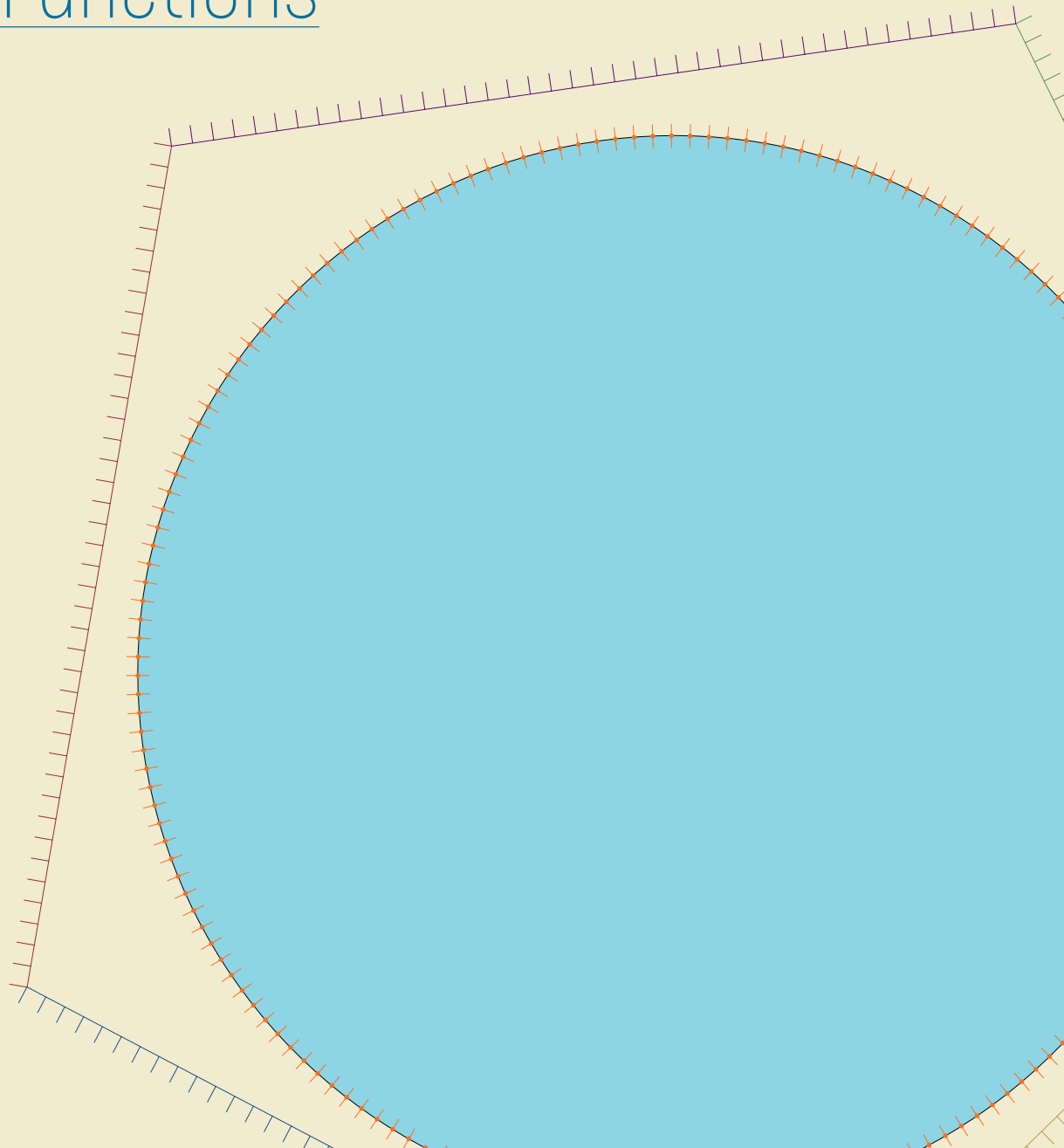
## Introduction

### **Background and Scope**

The use of artificial intelligence (AI) driven decision making in public functions has been touted around the world as a means of augmenting human capacities, removing bureaucratic fetters, and benefiting society. Yet, with concerns over bias, fairness, and a lack of algorithmic accountability, it is being increasingly recognized that algorithms have the potential to exacerbate entrenched structural inequality and threaten core constitutional values. While these concerns are applicable to both the private and public sector, this paper focuses on recommendations for public sector use, as standards of comparative constitutional law dictate that the state must abide by the full scope of fundamental rights articulated both in municipal and international law. For example, as per Article 13 of the Indian Constitution, whenever the government is exercising a "public function", it is bound by the entire range of fundamental rights articulated in Part III of the Constitution.

However, the definition and scope of "public function" is yet to be clearly defined in any jurisdiction, and certainly has no uniformity across countries. This poses a unique challenge to the regulation of AI projects in emerging economies. Due to a lack of government capacity to implement these projects in their entirety, many private sector organizations are involved in functions which were traditionally identified in India as public functions, such as policing, education, and banking. The extent of their role in any public sector project poses a set of important regulatory questions: to what extent can the state delegate the implementation of AI in public functions to the private sector?; and to what extent and how can both state and private sector actors be held accountable in such cases?.

AI-driven solutions are never "one-size-fits-all" and exist in symbiosis with the socio-economic context in which they are devised and implemented. As such, it is difficult to create a single overarching regulatory framework for the development and use of AI in any country, especially in countries with diverse socio-economic demographics like India. Configuring the appropriate regulatory framework for AI correctly is important. Heavy-handed regulation or regulatory uncertainty might act as a disincentive for innovation due to compliance fatigue or fear of liability. Similarly, regulatory laxity or forbearance might result in the dilution of safeguards, resulting in a violation of constitutional rights and human dignity. Therefore, we have sought to conceptualize optimal regulatory interventions based on key constitutional values and human rights that the state should seek to protect when creating a regulatory framework for AI. To devise these interventions, we identify a decision-making framework consisting of a set of core questions that can be used to determine the extent of regulatory intervention required to protect these values and rights.

We have arrived at the framework by identifying key values and rights, and analyzing AI use cases to understand how different uses and configurations of AI can challenge these values and rights. Specifically, the paper examines:

1. Use of AI in predictive policing by law enforcement;
2. Use of AI in credit rating by means of establishment of the Public Credit Registry (PCR) in India; and
3. Use of AI in improving crop yields for farmers.

This paper is divided into three sections. In the first section, we look at various models of regulation. In the second section, we expand on the use cases we chose to study in detail and the policy. In the third section, we identify core constitutional values that any regulatory framework on AI in the public sector should look to protect. In this section, we also highlight key regulatory interventions that need to be made to protect these values by developing a set of guiding questions.

We chose to work on the Indian ecosystem for three substantive reasons, apart from the convenience of geographic proximity, which allowed us to conduct our primary research. First, in terms of public policy advancement, we feel that working in India is important, as the technology and its governance frameworks are both in their nascent stages and the potential for the use of these technologies and their impact on the populace, especially those emerging technologically, is immense. Second, the constitutional framework in India on key issues such as privacy, discrimination, and exclusion has both a legacy of jurisprudence and is at a critical juncture, as they evolve and adapt with respect to emerging technologies. Finally, we believe that focusing on India allows us to make a unique contribution to the existing literature, as it charts out a potential regulatory model for other similarly placed emerging economies.

Our framework limits itself to decision making by a regulator when designing or deploying the AI solution. It does not delve into the adaptive regulatory strategy that needs to be devised as the AI project is implemented. It is also not an exhaustive framework, as many context-specific questions will alter its application. The objective is limited to framing broad questions that can guide specific regulatory interventions as decision makers choose to adopt AI.

## Methodology

From the outset, we realize that the term "artificial intelligence" is used in multiple ways, and its definition is often contested. For the purposes of this paper, we define AI as a dynamic learning system where a certain level of decision-making power is being delegated to the machine (Basu & Hickok, 2018). In doing so, we distinguish AI from automation, where a machine is being made to perform a repetitive task.

The first stage of our research involved studying three applications of AI in public functions. Through primary interviews and desk research, we sought to understand:

- How the decision was arrived at to devise an AI-based solution
- Relevant policy or political enablers or detractors
- What preparatory research or field work was done before implementing the solution
- How the data was gathered and collected
- Impact assessment frameworks or evaluation metrics used to determine the success of the project by the developers and implementers
- External assessments of the impact

Using what was learnt from these case studies, we created a decision-making framework that relied on key threshold questions, as well as possible regulatory tools that could be applied.

# Section I: Regulatory Models for AI

## Privatizing Public Functions

Across the world, activities traditionally undertaken by the state, including running prisons, policing, solving disputes, and providing housing and health services, are increasingly being delegated to private actors, often either private firms operating transnationally (Palmer, 2008) or quasi-governmental actors (Scott, 2017a). It is not only a shift in the extent of legislative discretion but the creation of formal and rule-based arrangements that were not needed in the welfare state model, where the state delivered all services directly (Scott, 2017a). Braithwaite's conception of a regulatory state combines state oversight with the commodification of service provision, where the citizen is treated as a consumer (Braithwaite, 2000). Businesses must deliver services with state oversight, but the extent of oversight and the modes of regulation must be determined contextually (Scott, 2017b).

The increasing privatization of public functions throws up two key constitutional questions. First, to what extent can public functions be delegated to a private actor? Little jurisprudence exists on this, as there have been very few challenges to privatization across jurisdictions. The Indian Supreme Court in *Nandini Sundar and Ors vs State of Chattisgarh* (2011), which banned the state designated private police organization, Salwa Judum, held that "modern constitutionalism posits that no wielder of power should be allowed to claim the right to perpetuate state's violence… unchecked by law, and notions of

innate human dignity of every individual" (Sundar and Ors v State of Chattisgarh, 2011). The Court went on to criticize the state of Chattisgarh's "policy of privatization" that was the cause of income disparity and non-allocation of adequate financial resources in the region, which in turn was responsible for the Maoist/Naxalite insurgency. However, there was no clarification on what services are "governmental" and cannot be delegated. The only clear carve out was the state's monopoly on the use of violence, which could under no circumstances be delegated. Although some indication of where to draw the line comes from the following dictum of the Supreme Court in *Nandini Sundar*:

> *"Policies of rapid exploitation of resources by the private sector, without credible commitments to equitable distribution of benefits and costs, and environmental sustainability, are necessarily violative of principles that are "fundamental to governance", and when such a violation occurs on a large scale, they necessarily also eviscerate the promise of equality before law, and equal protection of the laws, promised by Article 14, and the dignity of life assured by Article 21."*

The Israeli Supreme Court in *Academic Center of Law and Business vs Minister of Finance* (2006) had also invalidated a statute allowing for the privatization of prisons by reading its Basic Law. The judges in the majority opinion did not embark on an inquiry into whether private prisons worked better than those run by the government (Academic Center of Law and Business v Minister of Finance, 2006). Instead, there was an assumption made that privatization was illegal because private actors inherently harmed human rights more than public providers.[1] The Court argued that only the state itself had the right to deprive people of their liberty and dignity. The minority opinion countered this proposition by claiming that if the private sector was in fact able to maintain better prison conditions than the public sector, then privatizing prisons may actually further

human dignity instead of undermine it.[2] This is a valid concern for emerging economies as there are various circumstances, including AI deployment, where private actors can deliver services more efficiently than an overstretched state. However, given the implications for human rights and dignity, it is conceptually difficult to draw an objective line on delegation. The Court "assumed there is no constitutional impediment to privatization of a vast majority of services provided by the state".[3]

In the US, no bar to privatization exists and the market for private actors providing prison services is booming (Pelaez, 2019). In fact, a US appellate body judge has stated that a prisoner only "had a legally protected interest in the conduct of his keeper, not in the keeper's identity" (Pischke v Litscher, 1999). This lack of clarity on the definition, scope, and delegation of public functions means that when deciding the extent to which an AI use case can be delegated to a private actor, a number of other context-specific factors must be considered. These will be developed and discussed in Section III.

The second constitutional question hinges on the extent to which the state or a private actor can be held accountable for a violation of fundamental rights. The state action doctrine in the US formulates an apparently clear principle: constitutional rights apply to the state and not to private action (except in certain situations, such as *Habeas Corpus*).[4] State action, simply put, includes all government action which includes acts by the executive, legislature, and judiciary at both the central and state levels (Jaggi, 2017). However, the doctrine has a clear "public function" exception. As per this exception, a private actor may be considered a state actor if it "performs the customary functions of government" (Lloyd Corp Ltd v Tanner, 1972) or if it performs a function that is "traditionally exclusively reserved to the state" (Barrows v Jackson, 1953). The Indian Constitution is similar in that Article 12 states:

---

1. Para. 18 (Procaccia)

2. Id. ¶¶ 2, 4

3. Id. ¶ 65 (Beinisch) However, Justice Jowell did note that policing, defence, treaty-making, prosecution, and dissolving Parliament may be core governmental powers. ¶¶ 29–30

4. First articulated in The Civil Rights Cases (1883)

*"Definition in this part, unless the context otherwise requires, the State includes the Government and Parliament of India and the Government and the Legislature of each of the States and all local or other authorities within the territory of India or under the control of the Government of India."*

The question of whether private actors performing "public functions" comes under "other authorities" has come up before the Supreme Court. Questions have revolved around the status of the Board of Control for Cricket in India (BCCI). In *Zee Telefilms vs Union of India* (2005), the Supreme Court held that the BCCI is not discharging a public function, although it did not reject the public function test. The dissenting judges in *Zee Telefilms vs Union of India* (2005) recognized that with privatization and liberalization, as governmental functions are being delegated to private bodies, these private bodies must safeguard fundamental rights when discharging public functions. In 2015, the Supreme Court held that the BCCI is, in fact, performing a public function and therefore can be held accountable under Article 12 (Sethia, 2015). More recently, the Supreme Court held that a private university can be held accountable for violation of fundamental rights, as they are performing a public function or public duty by imparting education (Francis Coralie Mullin v UT of Delhi, 1981). Therefore, it is fair to say that Indian courts have adopted the public function exemption. Yet, given the lack of clarity on the definition of "public function", a context-specific approach is needed when ensuring that appropriate accountability, grievance redressal mechanisms, and liability are imposed in such cases. One test we recommend for the purpose of classification is linking the public function back to recognize aspects of the "right to life" enshrined in Article 21 of the Indian Constitution. The Supreme Court has held that "the right to life includes the right to live with human dignity and all that goes along with it, namely, the bare necessities of life such as adequate nutrition, clothing and shelter, and facilities for reading, writing,

and expressing oneself in diverse forms, freely moving about, and mixing and commingling with fellow human beings" (Francis Coralie Mullin v UT of Delhi, 1981). While recognizing that the magnitude and scope of this right is contingent on economic development, the Court stressed that the basic necessities of life, and the right to carry on such functions, are essential for basic human autonomy. Therefore, any entity carrying out a function that has implications for any of the functions described could be treated as a "public function", although this cannot operate as a hard and fast rule.

## Challenges to Regulating AI

Regulation is often designed to avert, mitigate, or limit risks (Haines, 2017) to human health or safety, or more broadly, to the effective functioning of a society. However, the risks that AI pose are only just being discovered and will continue to be realized as a greater number of use cases are designed and implemented. Importantly, the risks posed by AI cannot be determined only by evaluating the technology at hand. A genuine assessment of risk must contextualize the technology within the socio-economic, cultural, and demographic space within which it is being applied. The same AI technology or solution used for a specific use case in the defense industry may pose very different risks when used in the educational sector.

Scherer charts out four problems with regulating AI development ex ante (Scherer, 2016): "discreetness", which means that AI projects could be developed in the absence of large-scale institutional frameworks; "diffuseness", which entails that AI projects could be devised by a number of diffuse actors in various parts of the world; "discreteness", which means that projects will use discrete components and the final potential or risk of the AI system may not be apparent until the system finally comes together; and "opacity", which means that the technologies underpinning the system may be opaque to most regulators (Scherer, 2016).

Given these challenges, several academics have advocated applying Ayres and Braithwaite's proposition of responsive regulation to AI development (Terry, 2019). Simply put, responsive regulation suggests that appropriate regulatory interventions should be determined based on the regulatory environment and the conduct of the regulated (Ayres & Braithwaite, 1992).The crux of the idea lies in a pyramid of enforcement measures with the most interventionist command and control regulations at the apex and less intrusive measures such as self-regulation making up the base (Ayres & Braithwaite, 1992). For all matters, Ayres and Braithwaite believe it is better to start at the bottom of the pyramid and escalate up the structure if the regulatory objectives are not being met. This way, the government signals a willingness to regulate more intrusively while averting the negative impacts of more interventionist regulation at the very outset (Ayres & Braithwaite, 1992).

However, when deploying AI in public functions, moving from a spectrum of leniency to intrusiveness in all instances is fraught with risks to core constitutional values and human rights. This holds particularly true when the project is in its design stage or just about to be implemented, and the impact is not entirely known. We therefore advocate for "smart regulation" – a notion of regulatory pluralism that fosters flexible and innovative regulatory frameworks by using multiple policy instruments, strategies, techniques, and opportunities to complement each other (Gunningham & Sinclair, 2017). Based on certain threshold questions that help identify risks posed by a specific use case to core values, we attempt to provide guidance as to what different instruments, strategies, techniques, and opportunities could mitigate these risks associated with AI development and use.

## Modes of Regulation

Broadly speaking, "regulation" can be conceptualized as governing with a certain intention across a number of often-complex situations (Doekler,

2010) where competing interests are at stake (Kleinsteuber, n.d.). Traditionally, regulation has been determined by the sovereign, although market actors are increasingly determining their own regulatory frameworks, either through self-devised codes of conduct or in conjunction with sovereign entities. The decentralization of regulation away from a solely government-driven model is being spurred on by the fact that governments have incomplete information and expertise, and do not have the financial or human resources to devise, implement, and enforce regulation when emerging technologies propel rapid change and consequent uncertainty (Guihot, Matthew, & Suzor, 2017).

## Primary (Government-driven) Regulation

Traditionally, governments have various tools at their disposal to implement legislation. This includes nodality, authority, funding, and organization (Hood & Margetts, 2008). Nodality refers to the government's pivotal role as a receiver and distributor of vast sources of information, which enable it to ensure implementation of the law by detecting breaches and subsequently passing sanctions (Hood & Margetts, 2008). Authority bestows the government with the power to enforce sanctions and "demand, forbid, guarantee, and adjudicate" in a manner that is respected by all stakeholders (Hood & Margetts, 2008). In governmental regulation, implementation is through force and punitive sanctions for non-compliance, with the regulated not necessarily having a clear say in the framing of the regulation (Doekler, 2010). The treasure chest refers to the variety of resources, both monetary and infrastructural, at the disposal of the government to carry out any task (Hood & Margetts, 2008). Organization is the bureaucratic structure which enables the government to actualize the three other unique elements.

However, all of these elements may not necessarily apply to the multifarious nature of tasks that need to be examined when regulating AI-driven

solutions, particularly in economies as diverse and heterogeneous as India. The challenges in keeping up with the rapid pace of technological evolution have been better understood by private companies such as Google and Microsoft, who have taken the lead both in bank-rolling and implementing a variety of AI-driven solutions (Basu & Hickok, 2018). They possess the requisite expertise and human resources to conceptualize and incorporate various tools of regulation into the governance of AI. Therefore, in the regulatory domain, these companies are driving the rules of the game by creating codes of conduct for themselves and their peers in industry.

## Peer Regulation or Self-regulation

Jessop describes self-regulation as a system of bottom-up governance that allows private actors to limit the role of regulatory bodies by adopting a "reflexive self-organization of independent actors involved in complex relations of reciprocal interdependence, with such self-organization being based on continuing dialogue and resource sharing to develop mutually beneficial joint projects, and to manage the contradictions and dilemmas inevitably involved in such situations" (Jessop, 2003). In a self-regulatory ecosystem, actors conceptualize and voluntarily comply with their own set of codes, thereby serving as a form of informal regulation, with no punitive sanction for non-compliance (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020). Self-regulation can be one of two types. The first, more standardized form, describes situations where industry-wide organizations set rules, standards, and codes for all actors operating in that industry. The second, voluntarism, occurs when an individual firm chooses to regulate itself and create its own code of conduct without any coercion (Gunningham & Sinclair, 2017).

Attempts at self-regulation have already started in the governance of AI. A recent study (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020) by the Berkman-Klein Center at Harvard University (hereinafter the

"Berkman-Klein study") identified eight sets of "ethical" AI principles put forward by a range of multi-national companies, including Microsoft, Google (Pichai, 2018), and IBM. Each of these sets of guidelines espouse a set of principles that defer but fail to explicitly incorporate standards of domestic or international law (Basu & Pranav, 2019). For example, to protect the Right to Equality, the Google AI principles merely seek to avoid "unjust impacts on people, particularly to those related to sensitive characteristics", without referring explicitly to the various contours of and jurisprudence related to the Right to Equality across jurisdictions.

As identified by the European Commission High Level Expert Group, even after legal frameworks have been complied with, "ethical reflection can help us understand how the development, deployment, and use of AI systems may implicate fundamental rights and their underlying values, and can help provide more fine-grained guidance when seeking to identify what we should do rather than what we (currently) can do with technology" (European Commission, n.d.). However, Mittelstadt argues that ethical frameworks are prone to fail to regulate AI solutions because unlike other fields where ethics are used as regulatory interventions, AI lacks (1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms (Mittelstadt, 2019). Further, ethical guidelines devised by multi-national corporations often do not apply in the specific societal or legal contexts across jurisdictions (Arun, 2019).

Therefore, reliance on self-regulation through ethical AI guidelines may not be adequate to appropriately regulate the variety of ways in which AI may be deployed-in public functions and to genuinely protect core values and human rights.

## Co-regulation

A decentralized understanding of regulation entails an acknowledgement of the fact that states cannot be the only regulators, and the complexity, fragmentation, and the clashes in power and control ensure that regulation is hybrid, multi-faceted, and often indirect (Black, 2001). Co-regulation has a variety of definitions. Often referred to as "regulated self-regulation" (Schulz & Held, 2001), co-regulation is founded on a legal framework through which private entities govern their affairs through codes of conduct or set of rules (Doekler, 2010). The formation of the legal framework can be done in a multitude of ways but generally considers a link between state and non-state regulation. The European Commission has arrived at the following elements of co-regulation (Schulz & Thorsten, 2006):

1. The system is created to attain public policy objectives directed at societal processes;
2. There is a connection between the state and non-state regulatory system;
3. Some level of discretionary power is left to the non-state regulatory system;
4. There is an adequate level of supervision and involvement by the state.

In a co-regulatory framework, governments and private actors share responsibilities (Schulz & Thorsten, 2006). One way of doing this would be to divide up tasks. Government could set the high-level goals but enable the industry to set standards while still retaining some supervisory discretion.

Co-regulation is widely present in the US. For example, the Network Advertising Initiative (NAI) runs as a self-regulatory body that is then approved by the Federal Trade Commission (Federal Trade Commission Staff, 2009). Another form of co-regulation is when the government and private sector perform a number of tasks together. This may include both creation and enforcement of standards, such as in the case of the California Occupational Health and Safety Administration, which created a program where it worked with representatives from both management and labor to create and implement safety standards for construction sites (Freeman, 2000).

Through discussion and feedback, co-regulation would see the fostering of effective ideas over a period of time. A co-regulation approach to developing and implementing tools in AI governance would allow for the symbiosis of the private sector and technical expertise with the public sector and law-making experience. The potential problem with co-regulation is the creation of a culture of continuous lobbying, through which an already stretched public sector is compelled to respond to various pressure groups with conflicting agendas.

As we move from hierarchical regulation to more hands-off self-regulation, regulatory intervention becomes less rigid and binding, but also more participatory, and can potentially mitigate a far broader range of harms. Simply put, the greater the uncertainty and ambiguity in a type of intervention, the greater the range of cases it is able to regulate. The characteristics of each form of intervention have been summarized in the table below.

| Type of intervention | Enforceability | Rigidity | Creation | Applicability |
|---|---|---|---|---|
| **Legislation** | **Highest.** Binding law, along with clearly defined sanctions for non-compliance. | **Highest.** Clearly defined standards of municipal law with any ambiguity ideally being resolved by the judiciary. | **Top-down.** Devised by the legislator with optional consultation. | **Lowest common denominator.** Would only prevent directly identifiable harms resulting from AI. Would also require production of adequate evidence and causality. |
| **Co-Regulation** | **Middle.** Decentralized regulatory process may lead to a binding outcome. | Not unique. Could be clearly defined or vague depending on the outcome. | **Participatory.** With government, civil society and industry meaningfully engage in this process. | May have wide or narrow applicability to actors, situations, and individuals depending on the context. |
| **Self-Regulation** | **Lowest**. Enforceable at the organizational level but not binding. Reliance on "soft sanctions" with no clearly defined sanctions for non-compliance. | **Lowest**. Clearly articulated frameworks with greater ambiguity and more scope for manipulation. | **Participatory.** Devised through high-level consultations among industry and civil society but with an absence of government actors. | All AI that is ethical is necessarily legal. However, ethical frameworks have a broader applicability to harms that are outside the rigid confines of the law. |

**Table 1: Modes of regulation**

# Section II:
# Use Cases of AI in Public Functions

This chapter revolves around the governance of specific use cases that we studied concerning the use of AI in public functions in India. As the definition of "public function" remains unclear, we adopted a broad remit of use cases – from core governmental functions, which channel the state's monopoly over the use of violence (as discussed in Nandini Sundar), to credit rating, which is seeing increased private sector involvement and does not easily fit into the notion of a core state function such as lawmaking or policing.

The policy ecosystem in India has sought to promote AI adoption with a number of policy instruments, underscoring the need to instrumentalize AI and create broad stroke frameworks and focus areas. These include the discussion paper for the National Strategy on Artificial Intelligence, published by India's government think tank NITI AAYOG (Kumar, Shukla, Sharan, & Mahindru, 2018), as well as the Report of Task Force on Artificial Intelligence (Department for Promotion of Industry And Internal Trade, 2018) – a task force set up by the Ministry of Commerce. There are three main policy levers we can take away from the National Strategy. First, it suggests that the government should set up a multi-disciplinary committee to create a national data market place, so that organizations looking to derive data-driven insights can benefit from this data. Second, it proposes an "AI+X" approach that articulates the long-term policy vision for India. Instead of replacing existing processes in their entirety, decision making on AI should always look to identify a specific gap in an existing process (X) and add AI to augment efficiency. Third, it envisions the use of India as a garage bed for emerging economies, which we feel is a risky approach as it treats Indian citizens as guinea pigs without considering the potential impact on constitutional rights (Basu, 2019). Instead, India can set the tone for emerging economies by devising appropriate regulatory interventions that bring the best out of the technology without posing significant harms.

Without delving into the appropriate regulatory strategy for each use case, we explain each by looking at the following questions:

- How the decision was arrived at to devise an AI-based solution;
- Relevant policy or political enablers or detractors;
- What preparatory research or field work was done before implementing the solution;
- How the data was gathered and collected;
- Impact assessment frameworks or evaluation metrics used to determine the success of the project by the developers and implementers;
- External assessments of the impact;
- Extent of involvement of the private sector;
- Regulatory framework in the sector.

## Predictive Policing in Government/Law Enforcement

Predictive policing is making great strides in various Indian states, including Delhi, Punjab, Uttar Pradesh, and Maharashtra. A brainchild of the Los Angeles Police Department, predictive policing is the use of analytical techniques such as machine learning to identify probable targets for intervention to prevent crime or to solve past crime through statistical predictions (Berg, 2014). Conventional approaches to predictive policing begin by using algorithms to analyze aggregated data sets to map locations where crimes are concentrated (hot spots). Police in Uttar Pradesh (Sharma S. , 2018) and Delhi (Das, 2017) have partnered with the Indian Space Research Organization (ISRO) in a memorandum of understanding (MoU) that allows ISRO's Advanced Data Processing Research Institute to map, visualize, and compile reports about crime-related incidents.

There are also major developments on the facial recognition front. The Punjab Police, in association with Gurugram-based start-up Staqu, has begun

implementing the Punjab Artificial Intelligence System (PAIS), which uses digitized criminal records and automated facial recognition to retrieve information on the suspected criminal (Desai, 2019). Staqu has worked with police in a number of other states, including Uttar Pradesh, Uttarakhand, and Rajasthan (Ganguly, 2020).

It is important to acknowledge that bias existed in policing well before data-driven decision making came into the picture. Studies conducted in several states point to a disproportionately high representation of minorities and vulnerable communities in prisons (Common Cause, 2018). Muslims in particular have been impacted by this trend and have also reported the highest rates of contact with the police among any community (17%) (Common Cause, 2018). Courts have often found that incarceration has taken place based on false implications, which highlights flaws in the decision-making processes adopted by the police (Common Cause, 2018). This causes potentially flawed feedback loops, where increased police presence in certain areas is also leading to more crime being detected, in turn, leading to further police surveillance.[5]

The thinking behind devising and implementing predictive policing systems appears to be trust in the improved accuracy that data-driven decision making can provide. One official is reported as saying that "the key to [predictive policing] is the massive data on previous crimes and how best our people are able to analyze and correlate them with the present crimes" (Sharma, 2017).

A detailed analysis of Delhi Police's predictive policing by Marda and Narayan, entitled Crime Mapping, Analysis, and Mapping Systems (CMAPS), is very useful in understanding how this data is collected (Marda & Narayan, 2020a). The source of the input data was through calls received by the Delhi Police

Dial 100 call center. Unfortunately, the input data at this level is often flawed. The call taker is expected to enter the details of the crime into the "PA 100 form", which records information received from the caller into one of 130 pre-determined categories, or into "miscellaneous" if it is too difficult to slot them in cleanly. If more than one crime is reported, such as purse snatching and murder, only the more grievous crime is recorded. This is then escalated to the "Green Diary", which is often at the mercy of the police officer recording the incident. Police officers commonly believe that complaints by women are usually false (Marda & Narayan, 2020b). Marda and Narayanan's study confirms that gathering this information has been selective and subjective. Among police officers there is "a general apathy towards individuals living in slums and more forgiving outlooks with respect to individuals living in posh parts" (Marda & Narayan, 2020c).

The systems are shrouded in opacity, with CMAPS being out of the remit of the Right to Information Act, and appear to lack standard operating procedures or grievance redressal mechanisms. There is no legislation, policy, or guidelines that regulate and guide the operation of these systems, and no framework for evaluation. Reports indicate that there was no preparatory work or empirical research undertaken by the police to identify how concerns raised by multiple studies in other parts of the world where predictive systems have been deployed might play out in India. As Marda and Narayanan point out, the greater number of calls from poorer parts of Delhi might not be indicative of a higher crime rate than the relatively richer areas, it could simply be a cry of desperation from vulnerable communities who do not have access to other governance institutions (Khanikar, 2018). Given the current state of data curation practices, data-driven decision making might not provide a fair or accurate outcome.

---

5. Insights gained from primary interview

While there has been considerable political excitement about the use of AI and machine learning in law enforcement over the last few years (Basu & Hickok, 2018), there has also been parallel discourse advocating a need for caution about the use of such techniques. This cautionary note is even more pronounced in the use of machine learning by the state for public functioning, particularly where it leads to decision-making that impacts individual rights and entitlements. The intended use of AI by law enforcement in India to infer individual affect and attitude, offers a ripe opportunity to consider the opacity of such techniques. Even though the framers of the constitution deliberately kept the words "due process of law" out of the Indian Constitution, subsequent years of jurisprudence have adopted versions of the US constitutional law doctrines of "procedural due process" and "substantive due process" within the meaning of "procedure established by law" under Article 21. In criminal law, statutes that define offences and prescribe punishments are considered "substantive", while others relate to matters of process are considered "procedural". It is now accepted law that a procedural law which deprives "personal liberty" has to be "fair, just, and reasonable, not fanciful, oppressive, or arbitrary" (Maneka Gandhi v Union of India, 1978). During investigations, as per the criminal procedure code, law enforcement officers can take certain actions on the basis of "reasonable suspicion" and "reasonable grounds".

In the life cycle of actions by law enforcement agencies and the courts, starting from the opening of an investigation, followed by arrest, trial, conviction, and sentencing, we see that as the individual gets subject to increasing incursions or sanctions by the state, it takes a higher standard of certainty about wrongdoing and a higher burden of proof. Actions taken by law enforcement agencies, such as surveillance or arrests based on the use of sentiment analysis would be subject to the standard of due process. However, there is no way to judicially examine the reasonableness of such an action if the process is not explainable.

The standard in the US law for search and seizure under the Fourth Amendment is also of "reasonable suspicion", and we can look at US jurisprudence around this term for guidance. This standard was defined as requiring law enforcement agencies to "be able to point to specific and articulable facts which, taken together with rational inferences from those facts, reasonably warrant that [actions]" (Terry v Ohio, 1968). In the case of informant tips, US jurisprudence considers an informant's veracity, reliability, and basis of knowledge as relevant factors (Illinois v Gates, 1983). The standard of "reasonable suspicion" under the Fourth Amendment protection is not met by all tips. For instance, anonymous tips need to be detailed, timely, and individualized (Alabama v White, 1990). The grounds of reasonable complaint and credible knowledge in Section 49 of the Code of Criminal Procedure in India speak to a similar expectation of reliability and basis of knowledge.[6] It has also been clearly held that "reasonable suspicion" is not the same as the subjective satisfaction of a law enforcement officer (Partap Singh (Dr) v Director of Enforcement, Foreign Exchange Regulation Act, 1985), and clearly requires a good faith element on the part of the law enforcement agency (State of Punjab v Balbir Singh, 1994). In the case of a reliance upon an algorithm to substitute the role of tips, it is therefore necessary that the legal standards which can test the reliability and basis of an algorithmic technique, its suitability to the context, and the relevance of the dataset in use are evolved. However, where these techniques are opaque, as Marda and Narayanan have demonstrated, would severely limit the capacity of both law enforcement agencies to make informed decisions, as well as the ability of the judiciary to examine their use. When a law enforcement officer relies on tips to arrive at a good faith understanding, there is a clear way for a reviewing officer or a judge to evaluate the nexus between the available facts, good faith understanding, and the decisions taken – this is the basis of the review. The same is not possible in the case of an opaque algorithmic tool.

---

6. Section 49 (1) (a) of the Code of Criminal Procedure states as follows: "When police may arrest without warrant. (1) Any police officer may without an order from a Magistrate and without a warrant, arrest any person (a) who has been concerned in any cognisable offence, or against whom a reasonable complaint has been made, or credible information has been received, or a reasonable suspicion exists, of his having been so concerned."

There are also significant issues with judicial and law enforcement application of due process laws in India. For instance, despite having laws on admissibility and strict legal standards on what evidence is admissible, these rules are often set aside.[7] Even more alarming is the legal position on warrantless arrests, where the courts have held that police officers are not accountable for the discretion of arriving at the conclusion of reasonable suspicion while conducting a search on a suspect.[8] The lack of these protections make it harder to hold police accountable for excessive or unlawful use of predictive policing methods. Laws such as the Unlawful Activities (Prevention) Act (UAPA) are notorious for placing wide and unaccountable discretionary powers in the hands of law enforcement agencies (Khaitan N., 2019). In the UAPA, for instance, the term "unlawful activities" includes "disclaiming" or "questioning" the territorial integrity of India, and causing "disaffection" against India. The egregiously broad wording of such provisions come close to not just criminalizing unlawful acts but also objectionable beliefs and thoughts. In this context, the derivation of likelihood of an individual to commit crime through an opaque and unreliable technique such as predictive policing posits key challenges for decision makers.

## Credit Rating

AI is being harnessed by lenders to calculate credit scores and develop credit profiles. With the use of AI algorithms that draw from various data entries, such as an individual's banking transactions, their past decisions, their spending and earning habits, familial history, and mobile data, firms can make fast credit decisions for typical and atypical applicants (ICICI Bank, 2020). For example, Loan Frame uses AI and machine learning to examine a borrower's profile and evaluate their creditworthiness (Loan Frame, 2020). Similarly, start-ups such as Lending Kart (2020) and Capital Float (2020) use AI to assess the creditworthiness of micro, small, and medium enterprises (MSMEs) to help reduce the risk of defaulting. Kaleidofin is another start-up that has

attempted to solve the many challenges of financial inclusion in rural and semi-rural areas. They have used algorithms to analyze a variety of data and "recommend a single, seamless package of insurance and investment solutions" (Randazzo, 2013).

Companies and public sector banks assert that using AI has enabled them to bolster financial inclusion by including those who lack a formal credit history (Vishav, 2019, as cited in Singh & Prasad, 2020). Flaws in credit rating have existed across countries for some time (Smith, 2018), with the creditworthiness of an individual being contingent on local social and cultural notions of who "ought" to get loans, rather than simple number crunching (Kar, 2018a). Known as redlining, these practices have had deleterious financial and social impacts on minorities, particularly the African-American community in the US (Pearson, 2017; Corbett-Davies et al., 2017).

In a detailed exposition of what she terms the "moral economy of credit" in West Bengal, Kar demonstrates that bias on conceptions of "credit-worthiness" are entrenched among loan-givers across micro finance institutions (MFIs) (Kar, 2018a). She argues that "capacity was invoked as an ethical judgment [by the loan officer] of a borrower's ability to repay a loan, and was understood not through a seemingly objective analysis of financial data but through repeated exchanges with the borrowers during the verification process" (Kar, 2018b). She identifies five categories of exclusion driven by loan officers at microfinance institutions: religion, caste, class, language barriers, and location. Discrimination is "inter-sectional" (Kar, 2018b). "A number of Muslim dominated neighborhoods in Kolkata are discriminated against both because of their religion and because they are non-Bengali – largely migrants from the central Indian states of Uttar Pradesh or Bihar" (Kar, 2018b). The lack of data on individuals operating on the margins of or outside the formal financial system, combined with these entrenched patterns of exclusion, has ignited enthusiasm for data-driven decision making in this field.

---

7. See Umesh Kumar vs State of AP (2013) 10 SCC 591 ("It is a settled legal proposition that even if a document is procured by improper or illegal means, there is no bar to its admissibility if it is relevant and its genuineness is proved. If the evidence is admissible, it does not matter how it has been obtained. However, as a matter of caution, the court in exercise of its discretion may disallow certain evidence in a criminal case if the strict rules of admissibility would operate unfairly against the accused.")
8. Section 165 of the Code of Criminal Procedure

Machine learning algorithms are trained on curated datasets often referred to as "training data". For the purposes of fintech lending, this could be datasets that contain information about people's behavior online, spending patterns, living conditions, and geolocation, etc. As mentioned above, some fintech companies in India have publicly acknowledged that the number of data points is often around 20,000 (Nag, 2016). Machine learning-enabled credit scoring works by collecting, identifying, and analyzing data that can be used as proxies, as mentioned above, for the three key questions in any credit-scoring model: a) identity, b) ability to replay, and c) willingness to repay (Capon, 1982). With the advent of big data and greater digitization and datafication of information, new data sources such as telecom data, utilities data, retailers and wholesale data, and government data are available. Traditionally, credit-scoring algorithms consider set categories of data, such as an individual's payment history, debt-to-credit ratio, length of credit history, new credit, and types of credit in use.

The Reserve Bank of India is in the process of establishing the Public Credit Registry (PCR) for India – a comprehensive database of verified and granular information that will create a "financial information infrastructure" for providing credit at a national level. Chugh and Raghavan (2019) identified five limitations in the functioning of the existing information infrastructure, which the PCR seeks to remedy. These include a lack of comprehensive data, fragmented information, dependence on self-disclosure by borrowers, authenticity of the data, dated information, and inefficiencies due to multiple reporting (Chugh & Raghavan, 2019). Speaking about the registry, Dr. Viral Acharya, Deputy Governor, explained that "in an emerging economy like India, it is always felt that the smaller entrepreneurs, mostly operating under the informal economy do not get enough credit as they are informationally opaque to their lenders" (FinDev Gateway, 2019).

With the introduction of new forms of data, the richness of data may theoretically increase the predictive power of the algorithm (Ranger, 2018). However, narratives on greater accuracy presume both the suitability of input data towards the desired output, as well as faith that past attributes or activities that are used as training data do not lead to unintended outcomes (Joshi, 2020). There have been concerns that a combination of a vast variety of data points and the correlations recommended by machine learning processes will produce discriminatory outcomes that are not apparent and cannot be scrutinized in a court of law (Langenbucher, 2020). When a model relies on generalizations reflected in the data, the final result for the individual will be determined by shared data on the relative group that the system assigns to them, rather than the specific circumstances of the individual (Barocas & Selbst, 2016). Algorithmic

credit scores can remove bias only as much as the data that fuels them. Often, an assessment of the assigned group is also flawed. The development of "risk profiles" for individuals by the car insurance industry is a useful example (Kahn, 2020). Data might indicate that accidents are more likely to take place in inner city areas where the roads are narrower. Racial and ethnic minorities tend to reside more in these areas, which effectively means that the data indicates that racial and ethnic minorities, writ large, are more likely to get into accidents. Software engineers are responsible for constructing the mined datasets, defining the parameters and designing the decision trees. Therefore, as Citrone and Pasquale put it, "the biases and values of system developers and software programmers are embedded into each and every step of development" (Citron & Pasquale, 2014).

The roll out of algorithmic credit rating in India must be preceded by studies that map the possible disparate impacts of this practice and avoid some of the adverse impacts that have been experienced in other countries. Some companies have started taking individual steps to conduct grassroots level efforts (Kaleidofin, n.d.), but a larger industry-wide effort that is supported and endorsed by the government would be useful given India's depth and diversity. The government also needs to ensure regulatory certainty, so that start-ups are cognizant of the legal ecosystem within which they are operating.

Credit rating in India is governed by the Credit Information Companies (Regulation) Act, 2005 and the regulations issued in 2006 (Government of India, 2006). The Credit Information Companies (Regulation) Act, 2005, defines credit information as any information relating to the amounts and nature of loans, nature of securities taken, guarantee furnished, or any other funding-based facility given by a credit institution that is used to determine the credit-worthiness of a borrower. Given the variety of data that can be analyzed using algorithms, the definition might need revisiting (Goudarzi, Hickok, & Sinha, 2018).

As per Regulation 9.5.5 of the Credit Information Companies Regulation, 2006, it is mandatory for a bank that has rejected a loan on the basis of a credit information company report to:

(1) Send the borrower a written rejection notice within 30 days of the decision, along with (2) the specific reasons for rejection and (3) a copy of the credit information report, as well as (4) the details of any credit information company that constructed the report. If the decision has been rendered by crunching data through algorithms, the results must be human scrutable to the extent that a coherent explanation can be provided.

## Improving Crop Yields for Farmers

There has been a variety of initiatives taken by the government, in collaboration with the large technology companies, to equip farmers with more accurate information on weather patterns and ideal sowing dates for the generation of optimal crop yields (Gurumurthy & Bharthur, 2019).

IBM's Internet of things (IoT) platform has been used in many states in collaboration with NITI AAYOG – the Indian government's development think tank. The technology uses a "data fusion" approach which aggregates remote sensing meteorological data from The Weather Company, which is affiliated with IBM, along with satellite and field data (NASSCOM, 2018). In the state of Andhra Pradesh, Microsoft has collaborated with ICRISAT to develop an AI sowing app powered by the Microsoft Cortana Intelligence Suite. It sends advisories to farmers, providing them with information on the optimal date to sow by sending them text messages on their phones in their native languages. The government of Karnataka has signed a MoU with Microsoft to use predictive analytics for the forecasting of commodity pricing (UN ESCAP, 2019).

Despite being critical to India's economic development, the Indian agricultural sector continues to face a vast array of challenges (Indian Express, 2018): Some of them are associated with labor and resources, including migration to urban areas, overuse of groundwater, access to viable and quality seeds, a lack of balance in the use of fertilizers, and storage; infrastructure, including a lack of access to reliable credit, marketplaces, and technologies such as the Internet; and information, including a lack of access to reliable information about weather, markets,

and pricing (Nayak, 2015). Due to the information asymmetry in price modelling and forecasting, as well as weather and sowing conditions, specifically in Karnataka, the agricultural sector is characterized by a combination of drought-prone regions and areas that receive abundant irrigation (Deshpande, 2002). Compared to other states, Karnataka distinctively comprises a disproportionately large share of drought-prone areas (Deshpande, 2002). Farmer distress in Karnataka typically arises out of stress factors such as uncertainty in climatic factors and crop-prices (Deshpande, 2002). These conditions often have induced farmers to take miscalculated steps that result in onerous debts and sheer inability to meet family requirements (Deshpande, 2002). In addition, a study conducted in 2002 by the Karnataka State Agricultural Prices Commission identified that a large section of farmers (71%) did not end up selling their yield through regulated markets (Chatterjee & Kapur, 2016). This was because of an acute lack of knowledge (8%) of regulated markets (Chatterjee & Kapur, 2016).

Data-driven decision-making was targeted both by the state government of Andhra Pradesh and Telangana to address this specific gap (UN ESCAP, 2019). The implementation of the MoU was initiated through the development of an AI sowing app powered by the Microsoft Cortana Intelligence Suite, reported on June 9, 2016 (Reddy, 2016). Cortana Intelligence helps increase value in data by converting it into readily actionable forms (Heerdt, n.d.). This facilitates the expedient availability of information in achieving innovative outcomes within the agricultural industry. Using this intelligence, the app was able to interface
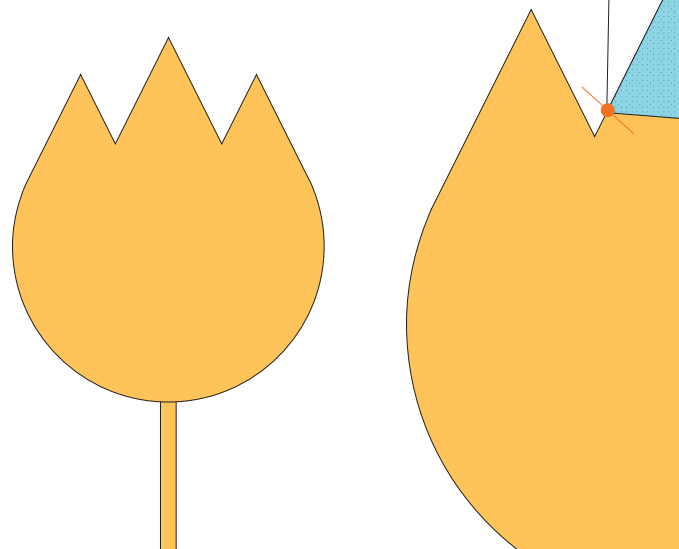
with models to forecast weather prepared by Where Inc. – a software company in the US. The app used extensive data mapping, including rainfall over the past 45 years in the Kurnool District (IANS, 2016; Reddy, 2016). The information was combined with data collected in the Andhra Pradesh Primary Sector Mission, popularly known as the Rythu Kosam Project (ICRISAT, n.d.). Launched with the objective of promoting productivity in the primary sector, the project involved the collection of household survey data relating, among other things, to crop yields (Charyulu, Shyam, Wani, & Raju, 2017). The combined data was downscaled in order to enable forecasting that could guide farmers in identifying the ideal week for the purpose of sowing (IANS, 2016).

The datasets considered relevant for the AI solution include yield-related information, weather, sowing area, and production. Part of the data was manually collected from farms in 13 districts in Karnataka by field officers deployed by ICRISAT during the aforementioned Rythu Kosam Project. The information was made available to Microsoft's Azure Cloud (Express Web Desk, 2017)  and subsequently downscaled to the village level in order to achieve the greatest possible precision, which was particularly useful for farmers in improving their decision-making capabilities. The machine learning software acquired by ICRISAT includes Cortana Intelligence and a personalized village advisory dashboard that uses business intelligence tools, both of which are prepared by Microsoft (ICRISAT, 2017).

In the pilot attempt implemented in Andhra Pradesh, the sowing period was estimated on the basis of datasets concerning the climate of the Devanakonda area in Andhra Pradesh, historically spanning a period of 30 years (1986–2015) (ICRISAT, 2017). The estimation involved computing data to forecast a future moisture adequacy index (MAI) based on data concerning daily rainfall, which was accumulated and reported by the AP State Development Planning Society (ICRISAT, 2017).

However, there were infrastructure-related hurdles to the successful implementation of both projects. As of December 2017, the overall Internet penetration in India was around 64.84% (20.26% in rural areas) (Agarwal, 2018).This meant that the AI intervention had to be very targeted. Since 77% of the bottom quintile owned a mobile phone (Bhattacharya, 2016), the output needed to be sent as text messages and not through an app that required the user to have a smart phone.

The NITI AAYOG reported that both in Karnataka and Andhra Pradesh there was an increase in crop yield between 10–30% due to the ICRISAT sowing advisory app (NITI Aayog, 2018). As a result of the MoU, the government can reportedly get price forecasts for essential commodities three months in advance in order to decide the minimum support price (IANS, 2017). The first impact assessment conducted in Devanakonda Mandal in Andhra Pradesh reflected a significant increase (30%) per hectare for farmers using the app (ICRISAT, n.d.). However, there are no publicly available reports on a holistic impact assessment of this project. Furthermore, the calculations undertaken to arrive at the 10–30% increase have also not been furnished.
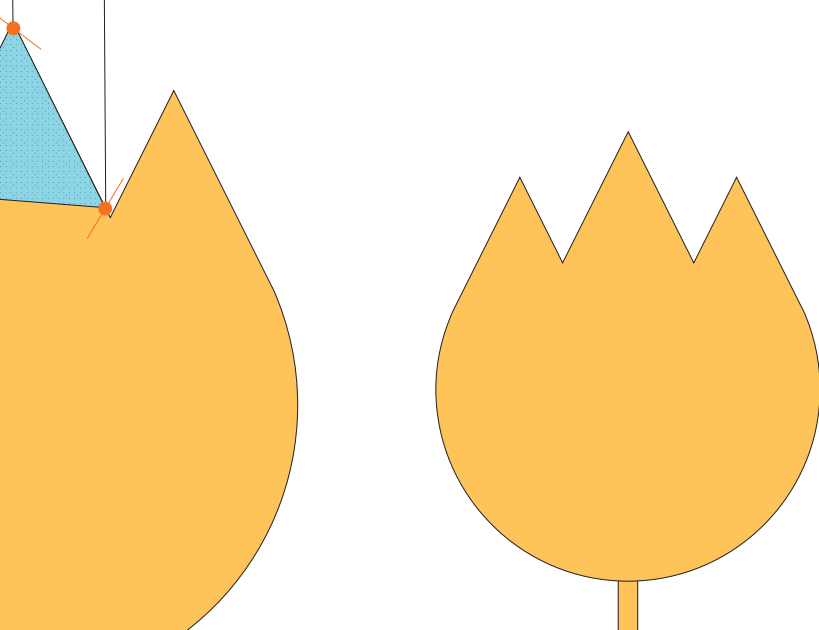
# Section III:
# Regulatory Interventions

To determine the optimal levels of regulation, we have arrived at a set of principles that enable the policymaker to define how the solution can work in consonance with existing values and constitutional frameworks as applicable to emerging economies. Transformative constitutionalism is a new brand of scholarship in comparative constitutional law, which celebrates the crucial role of the state and the judiciary in bringing about emancipatory change and rooting out structural inequality. Originally conceptualized as a Global South (Christiansen, 2011) concept designed as a counter-model to the individual rights-driven model of Northern Constitutions, scholars have now identified emancipatory provisions in several Western constitutions, such as Germany (Hailbronner, 2017). India's Constitution is one such example. The origins of constitutional order in India were designed to "bring the alien and powerful machine like that of the state under the control of human will" (Khilnani, 2004) and to eliminate the inequality of "status, facilities, and opportunities" (Kannabiran, 2012).

Therefore, a transformational approach necessarily considers the power asymmetries between the decision maker, implementer, and affected party, respectively. The questions for guiding regulation are an entry point that remedy the inherent asymmetries which span out in a variety of contexts.

As public authorities begin to adopt AI into decision-making processes for public functions, and begin to determine the ideal form of intervention(s), the extent to and the way in which decision-making capabilities can and are delegated to AI need to be questioned from the perspective of its transformative impact on justice, civil liberties, and human rights.

A framework of high-level articulation of values and guiding questions can help to guide these determinations. We curated the values based on an assessment both of India's constitutional ethos and an evaluation of values and rights that might inherently be tested by and therefore need to be explicitly protected when there is algorithmic decision making. This section contains an explanation of how we selected these questions and how they protect these values. It then goes on to draw out what an illustrative regulatory strategy might look like in response to these questions.

## Agency

Across jurisdictions, the concept of inherent dignity is connected to human agency – the capacity to make choices as one deems fit and pursue one's conception of a healthy life. Dignity reflected in agency does not require a specific set of criteria to define itself (Rao, 2013). It focuses on human capacities such as individuality, rationality, autonomy, and self-respect, and eschews focusing on the exercise of these traits (Rao, 2013).The Supreme Court of India has recognized the importance of the principle of autonomy in our constitutional schema and held that no discrimination by the state can undermine the personal autonomy of an individual (Bhatia, 2017).[9] Of the instruments demarcating ethical uses of AI, 69% have adopted a principle of human control. This essentially requires that key decisions delegated to AI remain under human review with a "human-in-the-loop" (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020).

Where stakeholders have sufficient agency to inform their use or interaction with AI, there is a presumption of limited regulatory intervention required. The less the agency of a stakeholder in dealing with AI, the greater the regulatory intervention needed.

---

9. Naz Foundation vs NCT of Delhi, (2009) 160 DLT 277 (High Court of Delhi). ("The grounds that are not specified in Article 15 but are analogous to those specified therein will be those which have the potential to impair the personal autonomy of an individual... Section 377 IPC in its application to sexual acts of consenting adults in privacy discriminates a section of people solely on the ground of their sexual orientation which is analogous to prohibited ground of sex."), see Tarunabh Khaitan, 'Reading Swaraj into Article 15: A New Deal for the Minorities' (2009) 2 NUJSLR 419

## Explanation

If adoption of an AI solution is mandatory, individual autonomy is immediately surrendered and the state determines the contours of individual agency. This is happening at present with the mandatory adoption of contact-tracing applications in light of the COVID-19 pandemic (Agrawal, 2020). During times of emergency or otherwise, if the state limits individual autonomy, then unique regulatory solutions that check the powers of the state must be deployed.

For AI solutions such as predictive policing, the primary users are state agents attempting to discharge their functions, whereas the impacted party is someone who is identified and evaluated by algorithmic decision-making. However, in the case of farmers receiving weather alerts, the farmer is both the primary user and the impacted party. To use another example, if the marketing and sales wing of a company uses sentiment analysis to analyze the user reviews of its products, the primary user, as well as the beneficiary or adversely impacted party of the analysis, is the company itself. On the other hand, if the same techniques are used for assessment of college application essays, the primary user is the university, but the parties who have to bear its adverse impact are the student applicants. Such a distinction must be made to determine if the potential risk of the algorithmic system is being borne by the stakeholders who choose to use it, or by other stakeholders who become unwitting victims of risks undertaken by others, and influences the impacted individual's ability to question the outcome or seek redress. Where parties choose to use systems marked by opacity and risk for commercial gains, there is a strong argument for regulatory restraint, unless the risks of such opaque decisions begin to percolate to others. In cases where the primary user and the impacted party are the same, there is a possibility for some opportunity for the user to play a role in deciding whether the inferences are used or not. In cases where they are not the same, the impacted party has no agency in this decision-making, and the further removed the role is, the potential for questioning this decision decreases when it is delegated to an algorithm.

## Questions

The following questions can help guide determinations of agency:

- Is the adoption of the solution mandatory?
- Does the solution allow for end-user control?
- What is the relationship between the primary user and impacted party?

## Recommended Regulatory Strategy

Adoption of the solution must be made mandatory only in exceptional circumstances. Compelling a farmer to adopt a technological solution constrains choice and undermines agency. Through primary regulation legislation or judicial decisions, we recommend that all states ensure that government entities at all levels adopt clear parameters for when any technological solution can be made mandatory. This must ensure that: (1) there is a pressing need in the public interest, (2) there is no reasonably available alternative, and (3) adequate measures of compensation, oversight, and grievance redressal are provided.

Even if the adoption of the solution is not mandatory, the power asymmetry between the user and impacted party needs to be closely considered. Where the power asymmetry is vast, such as police using AI to conduct surveillance in certain areas without the knowledge or consent of the people impacted, there needs to be far greater regulatory scrutiny. Ideally, this scrutiny should be multi-stakeholder and civil society groups, especially those representing vulnerable communities, and should be allowed to exercise vigilance by inputting into the design of the project before it is launched, auditing evaluation reports, engaging with targeted populations, and providing input as the project processes. Furthermore, training must be mandated for the public servants implementing the solution, thereby enabling them to understand the socio-economic complexities of those with whom they are engaging. Marda and Narayanan observed a lack of sensitization and empathy in the case of the Delhi police dealing with vulnerable communities (Marda & Narayan, 2020a) while Kar observed the same with loan officers passing judgement on "credit-

worthiness" (Kar, 2018a). Appropriate grievance redressal mechanisms that provide access for the vulnerable must be created. This should all be mandated through a top-down policy that is devised by the central government and made applicable to all government entities thinking of adopting AI solutions that have a great disparity between the end user and impacted party.

## Equality, Dignity, and Non-discrimination

### Background and Explanation

Human dignity is a core value recognized the world over, which the state should guarantee. In the Indian Constitution, dignity is mentioned in the Preamble and nowhere else. However, the Supreme Court has used the inclusion of the concept in the Preamble to interpret the guarantee of life and personal liberty to include a variety of traits associated with dignity. These include not only the bare necessities of life such as adequate nutrition, clothing, and shelter but also facilities for reading, writing, expressing oneself, and interacting with other human beings without fear (Mullin v And'r, Union Territory of Delhi, 1981).

When algorithms model and predict human behavior, there are important implications for the dignity of the individuals targeted. Modelling of human behavior includes use cases where the intent is either to predict or understand the activities, motivations, or proclivities of human beings. This is true even for cases where the intent is not to model human behavior but the clear implication is on decisions taken regarding human beings, due to systemic factors involved in data collection and labelling, use of algorithms, and impact of inferences, etc. As an individual's data is manipulated and formatted to extract a pattern about that individual's world, the individual or their data no longer exists for itself (Cheney-Lippold, 2017), but are massaged into various categories. Amoore terms this a "data-derivative", which is an abstract conglomeration of data that continuously shapes our futures (Amoore,

2011). Cheney-Lippold argues that algorithmic agents create identities for us on their own terms, rarely with input from the subjects of the algorithm itself (Cheney-Lippold, 2017) and terms this construction a measurable (a data equivalent of Weber's ideal type) construct of conceptual purity that does not occur in reality (Cheney-Lippold, 2017). Moreover, Rouvroy argues that the operation of the algorithm in terms of mathematical precision ignores the embodied individual and replaces him with a datafied substrate that can in no way capture the complexities of his character (Rouvroy, 2013). This leads to mathematical conclusions on the features of a certain group that might not reflect reality. Yet, the datafied substrate, replete with assumptions compounded by hidden layers, is used for making targeted decisions.

These ramifications are amplified in the case of minorities and other vulnerable communities. Algorithmic discrimination has been a concern among both legal experts and technologists for some time. Hao explains three phases at which some form of algorithmic bias might play out (Hao, 2019). The first stage comes with the framing of the problem. As soon as developers create a deep-learning model, they decide what output they want the model to provide and the rules needed to achieve this output. However, as discussed earlier, notions of "credit-worthiness", "recruitability", "suspicious", or "at risk" are often subject to cognitive bias. This makes it difficult to devise screening algorithms, which fairly portray society and the conglomeration of identities, and power asymmetries that define it (Basu, 2019).

The second stage is the data collection phase. As we saw with the predictive policing setup in Delhi, often data does not adequately represent reality. As crime rates are determined based on the number of calls that come into the Delhi Police call center, the quality of the dataset is highly dependent on how seriously the receiver takes each call (Marda & Narayan, 2020a). Calls from women from lower socio-economic groups

alleging sexual violence are often not taken seriously (Marda & Narayan, 2020a). A related problem is that datasets that are well curated and readily available are often very limited. For example, the data used for Natural Language Processing Systems for Parts of Speech (POS) tagging in the US come from popular newspapers such as The Wall Street Journal. However, accuracy of these datasets would decrease if the speech used by Wall Street Journal writers were applied to individuals or ethnic minorities who speak with a very different style (Blackwell, 2015).

The final stage is that of data preparation, where the developer selects the parameters which they want the algorithm to consider. For example, when determining credit-worthiness, the candidate's type of employment might be a parameter. It could be argued that someone working in the informal economy may be less likely to financially sustain themselves and thus would be deemed less credit-worthy. However, many individuals working in the informal economy in India are from lower caste communities (Kar, 2018a). Thus, working in the informal economy is an ostensibly neutral proxy for discriminating against a specific caste, thereby violating the right to equality when the data is being sorted during the machine learning process (Prince & Schwarcz, 2020).

The right to equality has been enshrined in several international human rights instruments and into the Equality Code of the Indian Constitution. The dominant approach to interpreting this right appears to focus on the grounds of discrimination in Article 15(1), thereby eschewing unintentional discrimination and disparate impact on certain communities. However, as Bhatia highlights (Bhatia, 2016), a few cases have considered indirect discrimination to some extent – an approach that is critical in the case of data-driven decision-making. Hence, we articulate the specific question on evaluating potential impact on minority groups, so that developers think of the potentially negative consequences of supposedly well-intentioned decisions.

## Guiding Questions

The following questions help guide regulations on agency, dignity, and non-discrimination:

- Is the AI solution modelling or predicting human behavior?
- Is the AI solution likely to impact individuals or communities, in particular the minority, protected, or at-risk groups?

## Recommended Regulatory Strategy

If AI is modelling or predicting human behavior, the state must be compelled to justify why this is necessary and proportionate to the objective. This justification must mandatorily be provided by any entity choosing to apply AI for this purpose, and must be enforced through either legislation or executive order. If a private sector actor such as Staqu is involved in partnership with the government, it must go through a process of accreditation, which should be determined by a co-regulatory body. All projects must also go through a mandatory impact assessment that considers the possibility of disparate impact or proxy discrimination. This must be mandated through co-regulatory guidelines framed by the government in consultation with private sector actors. We believe that a co-regulatory framework with regular consultations works best if a private sector actor is involved with the technology, as the government alone might not fully understand the implications of this technology. We also recommend that the private sector actor not be involved with the final decision. For instance, with credit rating, a number of private sector firms are involved in crunching data from the traditionally financially under-served and predicting their behavior. However, the final decision to sanction or reject a loan must be taken by a loan officer from a bank.

## Safety, Security, and Human Impact

The fundamental principle that guides regulatory decisions in this case is that of safety, security, and human impact. Where the use of AI has the potential for direct, adverse, or large-scale human impact, greater regulatory intervention is required. In the Berkman-Klein study, safety and security of AI systems are present in 81% of documents espousing ethical AI (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020). Therefore, the following broad questions need to be asked:

• Is there either a high likelihood or high severity of potential adverse human impact of the AI solution?
• Can the likelihood or severity of adverse impact be reasonably ascertained with existing scientific knowledge?

While we acknowledge that both likelihood and severity of impact, and the risks posed therein, are contextual, we believe that certain trends are worth noting. When AI systems model human behavior, it is much more likely to lead to an impact on the human beings in question, or those who may be seen as belonging to the same group or category by the algorithm. An AI solution that could cause greater harm if applied erroneously, such as one deployed for predictive policing, should be subject to more stringent standards, audits, and oversight than an AI solution designed to create a learning path for a student in the education sector. There could be cases where the behavior being modelled is not human, yet it could lead to significant human impact. For instance, an AI system that makes predictions about weather or environmental factors does not model human behavior but could be used to make assessments that directly impact human beings.

When considering the impact, it is imperative to look at both the severity and likelihood of the adverse impact. A high "likelihood" of harm indicates a high probability of the human rights, quality of life, and core value clusters being negatively impacted due to multiple pre-deployment factors, such as corrupted data sets or lack of awareness among users. Scale of harm indicates the extent of impact, which is determined by factors such as number of individuals impacted, while severity of harm can be determined by aspects such as clamping down on civil liberties or causing socio-economic distress.

In some cases, the likelihood of the adverse impact on human beings may be low, yet in the remote eventuality that it does lead to an adverse impact, its severity could be very high. For instance, the use of autopilot systems in aircraft navigation or in controlled trials where the number of people impacted are limited. The attention to both aspects of risk is essential, as often justifications for risky systems are based on low likelihood. However, even in cases where there is low likelihood of human harm, if the severity is high enough, it may still augur for greater regulatory scrutiny.

In situations where the likelihood or severity of harm cannot be reasonably ascertained, we recommend adopting the precautionary principle from environmental law and suggest that the solution not be implemented until scientific knowledge reaches a stage where it can reasonably be ascertained (Kriebel, et al., 2001).

### Regulatory Strategy

The following table contains a list of possible impact scenarios and regulatory strategies

| Outcome | Explanation Of Outcome | Recommended Regulatory Strategy |
|---|---|---|
| **A) High Likelihood, High Severity** | Scenarios where the state is involved in predicting human behavior (predictive policing/credit rating/predicting school dropouts) but training data is incomplete and a thorough impact assessment has not been conducted. | Ban or proscribe until underlying issues are solved to reduce likelihood of harm. If likelihood or severity cannot be gauged, then the solution must not be deployed. |
| **B) Low Likelihood, High Severity** | Scenarios where training data is robust but individuals relying on use case (flood prediction, crop price forecasting) may face dire economic consequences if solution works incorrectly. | State run human rights impact assessment that externally verifies compliance. |
| **C) High Likelihood, Low Severity** | Possible in pilot cases where data, methodology, and funding are not yet clear and safeguards have not been appropriately devised, or where AI is not directly impacting civil liberties or socio-economic rights (traffic management). | Strong redressal mechanisms that enable even one impacted individual to receive compensation, particularly if the initial estimation of severity is too low. |
| **D) Low Likelihood, Low Severity** | Where data is robust, methodology, troubleshooting, and outreach have been clearly devised, and use case is not directly impacting civil liberties or socio-economic rights. | Possible regulatory forbearance with strong industry-driven codes for standardization, evaluation, and redressal if private sector is involved. |

**Table 2: Impact thresholds**

## Accountability, Oversight, and Redress

### Background and Explanation

This principle attempts to grapple with two challenges to fostering accountability. The first challenge lies in the delegation of human decision making at some level to an algorithm, which creates an algorithmic "black box" through which inputs are processed and outputs are generated (Pasquale F., 2015). A certain level of transparency is key to fostering accountability frameworks for algorithmic decision-making. Any algorithmic decision-making framework in the public sector should reasonably be able to explain its decision to anyone impacted by its working. However, there may be a trade-off between the capacity or complexity of a model and the extent to which it can render a reasonably understandable explanation (Oswald, 2018).

Retrospective adequation is a legal standard we propose to promote algorithmic accountability (Sinha & Mathews, 2020). Essentially, this means that whenever inferences from machine learning algorithms influence decision making in public functions, they can do so only if a human agent is able to look at the existing data and discursively arrive at the same conclusion. Unlike the right to explanation under the General Data Protection Regulation (GDPR), which only includes "meaningful information about the logic involved, as well as the significance of the envisaged consequences of processing".[10] As opposed to the case of retrospective adequation, it does not tell us how an inference has been reached. This approach essentially draws from standards of due process and accountability evolved in administrative law, where decisions taken by public bodies must be supported by recorded justifications. Since the Maneka Gandhi vs Union of India judgment in 1978, the Supreme Court of India has clearly espoused the idea of both procedural and substantive procedural fairness. A further extension of this principle is the need for administrative authorities to record reasons to exclude or minimize arbitrariness (A Vedachalal Mudaliar v State of Madras, 1952). In some jurisdictions such as the UK and US, there are statutory obligations that require administrative authorities to give reasoned orders.[11] While there is no such corresponding statutory provision in India, the case law is fairly instructive in imposing similar obligations of quasi-judicial authorities (Travancore Rayons v Union of India, 1971; Siemen Engineering and Manufacturing Co. of India v Union of India, 1976). As Pasquale argues, explainability is important because reason-giving is intrinsic to the judicial process and cannot be jettisoned on account of algorithmic processing (Pasquale, F.A., 2017). The same principles equally apply to all administrative bodies, as it is a well-settled principle of administrative law that all decisions must be arrived at after a thorough application of mind. Much like a court of law, these decisions must be accompanied by reasons to qualify as a "speaking order". Where the administrative decisions are informed by an algorithmic process opaque enough to prevent this, the next logical question is whether a system can be built in such a way that it flags relevant information for independent human assessment to verify the machine's inferences. Only then will the requirements of what we call a speaking order be in any position to be satisfied.

Our assessment of opportunity for human supervision is based on the idea that where inferences are inherently opaque, they must provide sufficient information about the model and data analyzed, such that a human supervisor must be in a position to apply analogue

---

modes of analysis to the information available in order to conduct an independent assessment. For instance, where AI systems are used to detect hate speech for takedown from online platforms, it is possible to make available the inferences to a human supervisor who can apply her mind independently to the speech in question based on legal rules and standards on hate speech and relevant contextual information.

The increased role of the private sector in designing and deploying AI systems poses a challenge. As established earlier, there remains no clear threshold for demarcating public functions with private ones. With an increase in for-profit private actors playing a role in the discharge of functions that may be public, a liability mechanism that enables redress for adversely impacted individuals needs to be thought through. A potential thorny issue may be the proprietary nature of the source code, which the private sector developer may not want to share. This makes it imperative to think around unique regulatory interventions to constrain the private sector actor within the framework of the rule of law. This is particularly significant for start-ups, such as those involved in credit rating, who want to do "social good" but do not have the financial resources or bandwidth to create their own voluntary compliance strategy. Therefore, regulatory certainty that clearly demarcates scope of activity, liability, and evaluation metrics for private sector actors is vital.

The following questions help determine accountability, oversight, and redress:

• To what extent is the AI solution built with human-in-the-loop supervision prospects?
• Are there reliable means for retrospective adequation?
• Is the private sector partner involved with either the design of the AI solution, its deployment, or both?

## Smart Regulation Strategy

Since an empirical mapping of the potential loopholes in AI implementation across India's socio-economic demographics does not exist, all AI solutions must be built with human-in-the-loop supervision. Essentially, this means that while AI can aggregate and analyze data on a certain issue, the final decision will need to be taken by a human being. As our case studies showed, human bias in decision-making was prevalent well before machine learning came into the picture. However, human beings can be questioned, engaged with, and held accountable through legal proceedings – something that cannot be done with an AI system. In addition, human beings also retain the flexibility to make broader policy interventions. For example, if it is observed that crime rates are higher among a certain community, instead of merely trying to stamp out crime, a human being might try to identify the root cause of the crime, which might lie in higher rates of unemployment or poverty in the area. Therefore, they may look to intervene by devising social welfare programs instead of merely conducting enhanced surveillance. As such, human-in-the-loop must be made mandatory through top-down legislation.

Retrospective adequation is necessary for imposing accountability on AI systems discharging public functions and impacting citizens' rights. We recommend the evolution of technical standards from the private sector actors operating in India, which are then discussed and affirmed by a co-regulatory body such as the Bureau of Indian Standards.

If a private sector actor is involved with the design or deployment of the AI solution, then it must be first considered whether the activity in question falls within a reasonable and contextual understanding of a "public function". It is clear that private sector actors should

not deploy solutions when it comes to three core governmental functions: foreign relations, any form of violence or provision of security, and legislation. This essentially means that once the final decision is taken, any follow-up action must be decided and acted upon by a government entity.

Actors such as Staqu are involved in the design and development of the AI solution, even though the police implement the recommended outcome. Moreover, cases of public service delivery that have clear implications for the realization of the right to life could be considered public functions. Either the state or private actor must be held liable if rights are violated in the process. To encourage private actors to participate, the state may choose to soak up some of the liability for damages. However, clear mechanisms for assignment of liability must exist – something that was not done for Microsoft's partnership with the government of Karnataka. In such cases, consistent obligations must be imposed on the private sector. To this end, we recommend:

- Clearly drafted contracts with private sector developers that specify modes of liability, nature, and frequency of audits and impact assessments, as well as clarification that their source code and training data may need to be made public if the algorithmic decision-making is challenged in a court of law.
- Internal decision-making processes within the organization must be scrutinized for conformity with constitutional standards and human rights.
- The organization must ensure that they will not interfere with core government decision-making processes, such as deciding when to use violence in the interest of public order.

- In cases where private actors are involved with any function that violates civil and political or socio-economic rights, and an aggrieved individual(s) challenges the violation in a court of law, the court must treat this as a "public function" and hold the private sector actor to the same level of scrutiny as the government. If the government wants to shield the private actor from this liability, then it must be explicitly stated in the contract. These contracts must also be made public.
- That the private sector actor provides the needed capacity building to public sector actors to ensure they can understand the functioning and outputs of the system.

## Privacy and Data Protection

### Explanation
It is often argued that for emerging economies, the right to privacy should take a backseat to development. However, as we have highlighted in this paper, the poor and vulnerable are the most likely to have their civil liberties infringed by data-driven decision-making. When affirming the right to privacy as a fundamental right, the Indian Supreme Court strongly rebutted this, arguing that civil and political rights are important for every individual regardless of income (K. Puttaswamy v Union of India, 2017). They also affirmed that placing socio-economic rights over civil and political rights has been done away with by constitutional courts. Since this judgement in 2017, India has sought to formulate a data protection law – tabling a bill in Parliament in December 2019 (Basu & Sherman, 2020). While the obligations on private data processors in the bill are similar, it does some disservice to individual rights by granting the government a wide range of exceptions.

Section 35 states that exceptions can be made to collection rules, reporting requirements, and other requirements whenever the government feels that it is "necessary or expedient" in the "interests of sovereignty and integrity of India, national security, friendly relations with foreign states, and public order". The "necessary and expedient" standard replaces the "necessary and proportionate" standard laid down by the Puttaswamy judgement and reflected in a previous version of the bill tabled by the Justice B.N. SriKrishna Committee.

Another concern has been the bill's treatment of non-personal data (Basu & Sherman, 2020). Section 91(2) states that the government is allowed to direct data collectors to hand over anonymized personal information or other "non-personal data" for the purpose of "evidence-based policy making". Non-personal data is defined with little clarity as "anything that is not personal data". There has been a policy push towards channelizing as much data as possible towards social and economic development. The draft e-commerce policy defined data as "community data" to be owned and used for the benefit of all Indians (Government of India, 2019). On the other hand, chapter four of the Economic Survey treats data as a "public good", with no analysis of how this framework protects privacy rights. These concerns have been amplified as a result of the COVID-19 pandemic, where Indian citizens are being compelled to surrender personal data to the state through a contact-tracing app that has now become mandatory for download. Privacy is the most widely protected value across AI instruments – present in 97% of documents identified by the Berkman-Klein study (Government of India, 2019).

**Questions**
- Does the AI solution collect, use, and/or share personal data even in anonymized form?
- Can the identity of an individual be ascertained even if the system is not directly collecting or using personal information?

**Regulatory Strategy**

Whenever personal data is processed, there must be a national data protection law that demarcates user rights and redressal mechanisms in case of violations by both government and private sector actors. A specialized tribunal dealing with grievances under this law may be a co-regulatory, multi-stakeholder endeavor that has representatives from government, the private sector, and civil society. However, its decisions must be binding and enforced through primary, hierarchical legislation.

### Applying Regulatory Strategy to the Studied Use Cases

The following tables apply the regulatory strategies to the facts in the studied use cases. While not exhaustive, they indicate ways in which smart regulation that intervenes based on the guiding question can arrive at a comprehensive regulatory strategy that mitigates potential harms while enabling innovation. The regulatory interventions described in these tables are by no means an exhaustive framework that adequately tackles all systemic issues that some of these use cases may raise. Instead, they should act as illustrative guidelines that can guide policymakers to devise targeted interventions while simultaneously tackling larger societal questions and challenges through widespread structural changes.

<u>Regulatory interventions for predictive policing</u>

| Value | Questions | Predictive Policing | Regulatory Intervention |
|---|---|---|---|
| **Agency** | Is adoption of the solution mandatory? | Mandatory for all police officers depending on the decision made by police chief functionaries and mandatory for individuals that the police decide to use the solution on. | Regular consultation and feedback from all levels within the police hierarchy, in particular officers who directly engage with victims on the ground and the public.<br><br>Notice to individuals when a decision about them has been taken using an AI system.<br><br>Human rights impact assessment. |
| | Does the solution allow for end-user control? | Yes, as the police officer using it is the end user. | N/A |
| | Is there a vast disparity between the primary user and the impacted party? | Yes, between police officers and suspected criminals. | Mandatory certification for all police officers working both with the algorithm and implementing it on the ground (through notification).<br><br>Statistical standards for accuracy.<br><br>Evidentiary weight of decisions informed by an AI system. |

**Table 3a: Regulatory interventions for predictive policing**

| | | | |
|---|---|---|---|
| **Equality, Dignity, and Non-Discrimination** | Is the AI solution modelling or predicting human behavior? | Modelling criminality. | Needs assessment from the decision maker on why modelling human behavior is proportionate to the objective of reducing crime and also demonstrating why no other reasonable alternatives exist. |
| | Is the AI solution likely to impact minority, protected, or at-risk groups? | Possible disparate impact. | Awareness, sensitization, and creation of grievance redressal mechanisms and anti-discrimination regulations protecting vulnerable groups. |
| **Safety, Security, and Human Impact** | Is there a high likelihood or high severity of potential adverse human impact as a result of the AI solution? | Possible high likelihood and high severity, unless data collection practices are improved. | Proscription of solution until data curation and analysis is improved and standardized.<br><br>The use of the system should be guided by the principles of necessity, proportionality, and least intrusive means.<br><br>Compliance with international security standards. |
| | Can the likelihood or severity of adverse impact be reasonably ascertained with existing scientific knowledge? | Yes, through empirical research. | Government and the private sector should undertake regular empirical assessments of potential impact. |

**(Cont.) Table 3a: Regulatory interventions for predictive policing**

| | | | |
|---|---|---|---|
| **Accountability, Oversight, and Redress** | To what extent is the AI solution built with "human-in-the-loop" supervision prospects? | Human-in-the-loop exists. | |
| | Are there reliable means for retrospective adequation? | No publicly available information. | The private actor involved should mandatorily demonstrate possibility of retrospective adequation. |
| | Is the private sector partner involved with either the design of the AI solution, its deployment, or both? | Yes. | Contract as described above. Final implementation of the decision should continue to be done by the police. |
| **Privacy and Data Protection** | Does the AI solution use personal data, even in anonymized form? | Yes. | Any data collection must comply with a national data protection law that clearly separates personal and non-personal data. |

(Cont.) Table 3a: Regulatory interventions for predictive policing

<u>**Regulatory interventions for credit rating**</u>

| Value | Questions | Predictive Policing | Regulatory Intervention |
|---|---|---|---|
| **Agency** | Is adoption of the solution mandatory? | Optional for loan-providers from banks. They can potentially switch to a credit rating company that does not use AI. | Banks should have an internal regulatory strategy on the adoption of AI.<br><br>Human rights impact assessment. |
| | Does the solution allow for end-user control? | Yes, as the company/ bank engaging in credit rating is the end-user. | N/A |
| | Is there a vast disparity between the primary user and the impacted party? | Yes, there is a disparity between those generating the scores and those they are scoring. | Self-regulation: Loan officers and credit rating companies should communicate clearly to potential candidates the decision-making process, how AI is being used, and possible implications. |
| **Equality, Dignity, and Non-Discrimination** | Is the AI solution modelling or predicting human behavior? | It is determining "credit-worthiness". | Mandatory needs assessment from bank clarifying why algorithmic decision-making is more accurate than traditional credit scoring methods, as well as full transparency on data being used and curation methods. |
| | Is the AI solution likely to impact minority, protected, or at-risk groups? | Possible disparate impact. | Awareness, sensitization, training and creation of grievance redressal mechanisms targeting vulnerable groups. |

**Table 3b: Regulatory interventions for credit rating**

| | | | |
|---|---|---|---|
| **Safety, Security, and Human Impact** | Is there a high likelihood or high severity of potential adverse human impact as a result of the AI solution? | Possible high likelihood and high severity. | Mandatory pilot projects and standardization of data curation practices certified by a co-regulatory committee. |
| | Can the likelihood or severity of adverse impact be reasonably ascertained with existing scientific knowledge? | Yes. | |
| **Accountability, Oversight, and Redress** | To what extent is the AI solution built with "human-in-the-loop" supervision prospects? | Human-in-the-loop exists. | |
| | Are there reliable means for retrospective adequation? | No publicly available information. | Retrospective adequation should comply with Indian credit regulations. |
| | Is the private sector partner involved with either the design of the AI solution, its deployment, or both? | Both. | Contract as described above. If the private sector partner is a start-up, the state may choose to cushion some of the liability. Final decision must be independently taken by the bank sanctioning the loan. |
| **Privacy and Data Protection** | Does the AI solution use personal data, even in anonymized form? | Yes. | Any data collection must comply with a national data protection law that clearly separates personal and non-personal data. |

**(Cont.) Table 3b: Regulatory interventions for credit rating**

**Regulatory interventions for AI in agriculture**

| Value | Questions | Agriculture | Regulatory Intervention |
|---|---|---|---|
| **Agency** | Is adoption of the solution mandatory? | No, farmers may opt out. | Pros and cons of adopting the solution should be clearly communicated in an understandable format to the farmer (self-regulation). |
| | Does the solution allow for end-user control? | Yes, the farmer using the solution is the end-user. | N/A |
| | Is there a vast disparity between the primary user and the impacted party? | No, the farmer is the end-user and feels the impact of the solution. | A co-regulatory consultative body should be set up to organize regular consultations between the users and the developers of the project. |
| **Equality, Dignity, and Non-Discrimination** | Is the AI solution modelling or predicting human behavior? | It is modelling crop patterns and weather data. | |
| | Is the AI solution likely to impact minority, protected, or at-risk groups? | No, while there may be a negative impact, it is unlikely to specifically impact minorities. | All farmers may not equally benefit from the app. Government and private sector partners must mandatorily provide training, set up a pre-requisite infrastructure to the extent possible, and also study trends on why certain farmers may not be benefitting. |

**Table 3c: Regulatory interventions for AI in agriculture**

| | | | |
|---|---|---|---|
| **Safety, Security, and Human Impact** | Is there a high likelihood or high severity of potential adverse human impact as a result of the AI solution? | Depending on the quality of the data curated, there is possible low likelihood and low severity. | Mandatory pilot projects and standardization of data curation practices as certified by a co-regulatory committee. |
| | Can the likelihood or severity of adverse impact be reasonably ascertained with existing scientific knowledge? | Yes. | The private sector partner could publish research on preliminary scientific studies (voluntarism). |
| **Accountability, Oversight, and Redress** | To what extent is the AI solution built with "human-in-the-loop" supervision prospects? | Unclear. | More public information about the working of the app should be disclosed to the public and to the farmers concerned. |
| | Are there reliable means for retrospective adequation? | No publicly available information. | The private sector partner should be able to provide retrospective adequation for all decisions. |
| | Is the private sector partner involved with either the design of the AI solution, its deployment, or both? | Both. | There needs to be a contract clearly imposing liability on the private sector partner in case of negligence. If the private sector partner is a start-up, the state may choose to cushion some of the liability. |
| **Privacy and Data Protection** | Does the AI solution use personal data, even in anonymized form? | Yes. | Any data collection must comply with a national data protection law that clearly separates personal and non-personal data. |

**(Cont.) Table 3c: Regulatory interventions for AI in agriculture**

# Conclusion

The application of regulatory interventions to use cases brought up a number of similarities. While predictive policing is a core government function that could involve violence further down the line, the *modus operandi*, and therefore the potential threats to core constitutional values are similar to those in credit rating. The fundamental difference between these two use cases and the agricultural case study is that these involved two sets of human beings – one group being in a position of power that is attempting to predict how less powerful human beings will act. Thus, the regulatory interventions needed to optimally govern AI stem from those necessary to remedy structural injustices in society. The danger, however, in both India and other parts of the world, stems from technological solutionism, which assumes that existing societal fissures can be occluded through data-driven decision making. The reality is quite different, with data-driven decision making needing to adapt the same values that were required to fairly govern society in a pre-AI world. This is compounded by a lack of effective public oversight and consultation of both policymaking and technological implementation. There are no publicly scrutable external impact assessments post-deployment or publicly available empirical socio-economic assessments prior to deploying the solution.

Our paper establishes a framework for adapting these values through a series of questions that identify critical junctures at which core constitutional values and human rights may be at threat due to algorithmic decision-making. Our framework is by no means exhaustive and is meant to be read as a set of guidelines for decision makers and technologists looking to devise their own set of frameworks. The set of regulatory tools mapped out by Freiburg (2010) may remain relevant and need to be applied across contexts – often in response to knowledge that may be gained as the AI solution is implemented, evaluated, and adapted.

The five sets of values that we felt merited protection: (1) agency; (2) equality, dignity, and non-discrimination; (3) safety, security, and human impact; (4) accountability, oversight, and redress; and (5) privacy and data protection, were selected not only from a study of India's constitutional fiber but also through an assessment of AI policy instruments released by a variety of stakeholders around the world. As such, we feel that our framework – although researched and developed in an Indian context – applies across emerging economies who desire to improve the government's role in public service delivery while still mitigating negative impacts.

A core challenge continues to be the complex question of the involvement of the private sector in functions that have traditionally been the government's prerogative, and often those that have implications for fundamental rights. One of the most important recommendations of our paper centers around the need to hold the private sector accountable in these instances through uniformly worded contracts that adequately impose liability along with the delegation of any responsibility. However, given the lack of government capacity to entirely identify, design, and deploy an AI-driven solution, some regulatory room must be given to these actors to innovate.

Appropriate regulation therefore does not fit neatly into the division of the modes of hierarchical regulation, co-regulation, and self-regulation. A smart regulatory strategy would require a combination of all three.

Going forward, we feel the need for more empirical assessment of use cases in emerging economies, as much of the literature, both on the technology and regulatory frameworks, are devised in a Western context and therefore not entirely applicable to emerging economies. That said, our paper shows that algorithmic decision-making is becoming more commonplace in emerging economies. Through a close analysis of the information gained from these empirical assessments and a strong commitment to the values described, we believe that adequate ex ante regulation can mitigate harms while also enabling the realization of prospects for social good.

## Acknowledgements

## References

*A Vedachalal Mudaliar v State of Madras*, AIR Mad. 276 (1952)

*Academic Center of Law and Business v Minister of Finance*, Isr. (2006, Aug 20)

Agarwal, S. (2018, February 20). Internet users in India expected to reach 500 million by June: IAMAI. Retrieved from The Economic Times: https://economictimes.indiatimes.com/tech/internet/internet-users-in-india-expected-to-reach-500-million-by-june-iamai/articleshow/63000198.cms

Agrawal, A. (2020, May 1). Lockdown Extension: Aarogya Setu Mandatory for All Employees and in Containment Zones. Retrieved from MEDIANAMA: https://www.medianama.com/2020/05/223-coronavirus-lockdown-extended-by-2-weeks-country-divided-into-red-orange-and-green-zones/

*Alabama v White*, 496 U.S. 325 (1990)

Amoore, L. (2011). Data Derivatives: On the Emergence of a Security Risk Calculus for Our Times. SAGE journal, 24, 27. Retrieved from https://journals.sagepub.com/doi/10.1177/0263276411417430

Arun, C. (2019). AI and the Global South: Designing for Other Worlds. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), the Oxford Handbook of Ethics of AI. Oxford University Press. Retrieved from https://ssrn.com/abstract=3403010

Ayres, I., & Braithwaite, J. (1992). Responsive Regulation. Oxford University Press.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. 104 California Law Review 671.

*Barrows v Jackson*, 252, U.S. (1953)

Basu, A. (2019, October 12). We Need a Better AI Vision. Fountainink. Retrieved from Fountain Ink: https://fountainink.in/essay/we-need-a-better-ai-vision-

Basu, A., & Hickok, E. (2018). Artificial Intelligence in the Governance Sector in India. India: The Centre for Internet and Society. Retrieved from https://cis-india.org/internet-governance/ai-and-governance-case-study-pdf

Basu, A., & Pranav, M. (2019, July 21). What is the problem with 'Ethical AI'? An Indian Perspective . Retrieved from The Centre for Internet and Society: https://cis-india.org/internet-governance/blog/what-is-the-problem-with-2018ethical-ai2019-an-indian-per

Basu, A., & Sherman, J. (2020, January 23). Key Takeaways from India's Revised Personal Data Protection Bill. Lawfare.

Berg, N. (2014, June 25). Predicting Crime, LAPD style. Retrieved from The Guardian: https://www.theguardian.com/cities/2014/jun/25/predicting-crime-lapd-los-angeles-police-data-analysis-algorithm-minority-report

Bhatia, G. (2016). Retrieved from Indian Constitutional Law and Philosophy: https://indconlawphil.wordpress.com/tag/indirect-discrimination/

Bhatia, G. (2017). Equal moral membership: Naz Foundation and the refashioning of equality under a transformative constitution. Indian Law Review, 115-144.

Bhattacharya, P. (2016, December 5). 88% of households in India have a mobile phone. Retrieved from Livemint: https://www.livemint.com/Politics/kZ7j1NQf5614UvO6WURXfO/88-of-households-in-India-have-a-mobile-phone.html

Black, J. (2001). Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a 'Post-Regulatory' World 54 Current Legal Problems (Vol. 54). Current Legal Problems.

Blackwell, A. F. (2015). Interacting with an Inferred World: The Challenge of Machine Learning for Humane Computer Interaction". Proceedings of the Fifth Decennial Aarhus Conference on Critical Alternatives, 179.

Braithwaite, J. (2000). The New Regulatory State and the Transformation of Criminology. British Journal of Criminology, 40, 222-38. http://doi:10.1093/bjc/40.2.222

Capital Float (2020). Retrieved from Capital Float: https://capitalfloat.com/

Capon, N. (1982). Credit Scoring Systems: A Critical Analysis. Journal of Marketing, 46(2), 82-91. Retrieved from https://www.jstor.org/stable/3203343

Charyulu, D. K., Shyam, D. M., Wani, S. P., & Raju, K. (2017). Rythu Kosam: Andhra Pradesh Primary Sector Mission. Coastal Andhra Region Baseline Summary Report.

ICRISAT Development Center. Retrieved from http://111.93.2.168/idc/wp-content/uploads/2018/01/IDC-Report-No-13-Rythu-Kosam.pdf

Chatterjee, S., & Kapur, D. (2016). Understanding Price Variation in Agricultural Commodities in India: MSP, Government Procurement, and Agriculture Markets. India Policy Forum. Retrieved from http://www.ncaer.org/events/ipf-2016/IPF-2016-Paper-Chatterjee-Kapur.pdf

Cheney-Lippold, J. (2017). We Are Data: Algorithms and the Making of Our Digital Selves. NYU Press.

Christiansen, E. C. (2011, January 1). Transformative Constitutionalism in South Africa: Creative Uses of Constitutional Court Authority to Advance Substantive Justice. SSRN. Retrieved from https://ssrn.com/abstract=1890885

Chugh, B., & Raghavan, M. (2019, June 18). The RBI's proposed Public Credit Registry and its implications for the credit reporting system India. Retrieved from Dvara Research: https://www.dvara.com/blog/2019/06/18/the-rbis-proposed-public-credit-registry-and-its-implications-for-the-credit-reporting-system-in-india/

Citron, D. K., & Pasquale, F. A. (2014). The Scored Society: Due Process for Automated Predictions. Washington Law Review, 14, 89.

Commission, E. (n.d.). Ethics Guidelines for trustworthy AI. Retrieved from European Commission: https://ec.europa.eu/futurium/en/ai-alliance-consultation

Common Cause. (2018). Status of Policing in India Report 2018: A Study of Performance and Perceptions. Common Cause & Lokniti - Centre for the Study Developing Societies (CSDS). Retrieved from https://www.commoncause.in/pdf/SPIR-2018-c-v.pdf

Corbett-Davies, S. (2017). Algorithmic Decision-making and the Cost of Fairness. Stanford University. Retrieved from http://www.antoniocasella.eu/nume/Corbett-Davies_2017.pdf

Das, S. (2017, March 21). How Predictive Analytics Helps Indian Police Fight Crime. Retrieved from http://www.computerworld.in/feature/how-predictive-analytics-helps-indian-police-fight-crim

Department for Promotion of Industry and Internal Trade. (2018). Report of Task Force on Artificial Intelligence. Government of India. Retrieved from https://dipp.gov.in/whats-new/report-task-force-artificial-intelligence

Desai, K. (2019, March 31). Now Police Use Apps to Catch a Criminal. Retrieved from Times of India: https://timesofindia.indiatimes.com/home/sunday-times/now-police-use-apps-to-catch-a-criminal/articleshow/68649118.cms

Deshpande, R. S. (2002, June 29). Suicide by Farmers in Karnataka Agrarian Distress and Possible Alleviatory Steps. Economic and Political Weekly, pp. 2601-2604. Retrieved from http://shreeindia.info/rsdeshpande.com/wp-content/uploads/2014/03/Suicide_by_Farmers_in_Karnataka.pdf

Doekler, A. (2010). Self-regulation and Co-regulation: Prospects and Boundaries in an Online Environment. Master of Law thesis, University of British Columbia. Retrieved from https://open.library.ubc.ca/cIRcle/collections/ubctheses/24/items/1.0071207

Express Web Desk. (2017, October 27). Karnataka govt inks MoU with Microsoft to use Artificial Intelligence for digital agriculture. Retrieved from The Indian Express: https://indianexpress.com/article/india/karnataka-govt-inks-mou-with-microsoft-to-use-artificial-intelligence-for-digital-agriculture-4909470/

Federal Trade Commission Staff. (2009). Report on Self-regulatory Principles for Online Behavioral Advertising. Retrieved from https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-staff-report-self-regulatory-principles-online-beh

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020, January 15). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center Research. Retrieved from https://ssrn.com/abstract=3518482

Francis Coralie Mullin v UT of Delhi, AIR 746 (1981).

Freeman, J. (2000). The Private Role in Public Governance. NYULR, 75, 543, 547, 651–53.

Freiberg, A. (2010). Restocking the Regulatory Tool-kit. Dublin. Retrieved from http://www.regulation.upf.edu/dublin-10-papers/1I1.pdf

Ganguly, S. (2020, March 31). Gurugram-based Start-up Staqu Has Notified AI-powered JARVIS to Battle Coronavirus. Retrieved from Your Story: https://yourstory.com/2020/03/gurugram-ai-startup-staqu-jarvis-coronavirus

Gateway, F. (2019, January 29). India: Reserve Bank of India Is Working on Public Credit Registry to Improve Access to Micro Credit. Retrieved from FinDev Gateway: https://www.findevgateway.org/news/india-reserve-bank-india-working-public-credit-registry-improve-access-micro-credit

Goudarzi, S., Hickok, E., & Sinha, A. (2018). AI in Banking. India: The Centre for Internet and Society. Retrieved from https://cis-india.org/internet-governance/files/ai-in-banking-and-finance

Government of India. (2006). Notification. India. Retrieved from https://rbidocs.rbi.org.in/rdocs/Content/PDFs/69700.pdf

Government of India. (2019). Data "Of the People, By the People, For the People."

Government of India. (2019). Draft National E-Commerce Policy. Retrieved from https://dipp.gov.in/sites/default/files/DraftNational_e-commerce_Policy_23February2019.pdf

Guihot, M., Matthew, A. F., & Suzor, N. P. (2017). Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence. VJETL, 20(2), 385, 429. Retrieved from http://www.jetlaw.org/wp-content/uploads/2017/12/2_Guihot-Article_Final-Review-Complete_Approved.pdf

Gunningham, N., & Sinclair, D. (2017). Smart Regulation. In P. Drahos (Ed.), Regulatory Theory: Foundations and Applications (p. 115). ANU Press.

Gurumurthy, A., & Bharthur, D. (2019). Taking Stock of AI in Indian Agriculture. IT for Change. Retrieved from https://itforchange.net/sites/default/files/1664/Taking-Stock-of-AI-in-Indian-Agriculture.pdf

Hailbronner, M. (2017, November 22). Transformative Constitutionalism: Not Only in the Global South. American Journal of Comparative Law, 65(3), 527-556. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2777695

Haines, F. (2017). Regulation and Risk. In P. Drahos (Ed.), Regulatory Theory, Foundations and Applications (p.181). ANU Press.

Hao, K. (2019, February 4). This is how AI bias really happens—and why it's so hard to fix. Retrieved from MIT Technology Review: https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Heerdt, J. (n.d.). Transform your data into intelligent action with Cortana Analytics Suite. Retrieved from Sogeti: https://www.sogeti.nl/sites/default/files/Transform%20your%20data%20into%20intelligent%20action%20with%20Microsoft%20Cortana%20Analytics%20Platform.pdf

Hood, C. C., & Margetts, H. Z. (2008). The Tools of Government in the Digital Age. Palgrave Macmillan.

IANS. (2016, June 9). Microsoft develop sowing app for Andhra Pradesh farmers. Retrieved from Financial Express: https://www.financialexpress.com/industry/technology/microsoft-develop-sowing-app-for-andhra-pradesh-farmers/279171/

IANS. (2017, December 19). #GoodNews: Indian Farmers Go the AI Way to Increase Crop Yields. Retrieved from the quint: https://www.thequint.com/news/india/good-news-indian-farmers-use-ai-for-higher-crop-yields

ICICI Bank. (2020, January 1). Artificial Intelligence in Loan Assessment: How does it Work?' Retrieved from ICICI Bank: https://www.icicibank.com/blogs/personal-loan/artificial-intelligence-in-loan-assessment-how-does-it-work.page?

ICRISAT. (2017, January 9). Microsoft and ICRISAT's Intelligent Cloud Pilot for Agriculture in Andhra Pradesh Increase Crop Yield for Farmers. Retrieved from ICRISAT: http://www.icrisat.org/microsoft-and-icrisats-intelligent-cloud-pilot-for-agriculture-in-andhra-pradesh-increase-crop-yield-for-farmers/

ICRISAT. (2017, January 13). New Sowing Application Increases Yield by 30%. Retrieved from ICRISAT: http://www.icrisat.org/new-sowing-application-increases-yield-by-30/

ICRISAT. (n.d.). Microsoft CEO Speaks on Collaboration with ICRISAT. Retrieved from ICRISAT: http://www.icrisat.org/microsoft-ceo-speaks-on-collaboration-with-icrisat/

ICRISAT. (n.d.). Rythu Kosam. Retrieved from ICRISAT: http://www.icrisat.org/tag/rythu-kosam

*Illinois v Gates*, 462 U.S. 213 (1983)

Indian Express (2018, March 16). Why are India's Farmers Committing Suicide? Retrieved from Indian Express: http://www.newindianexpress.com/nation/2018/mar/15/why-are-indias-farmers-committing-suicide-1787539.html.

Jaggi, S. (2017). State Action Doctrine. Max Planck Encyclopedia of Comparative Constitutional Law. Retrieved from https://oxcon.ouplaw.com/view/10.1093/law-mpeccol/law-mpeccol-e473

Jessop, R. (2003). Governance and Metagovernance: On Reflexivity, Requisite Variety, and Requisite Irony. Sociology. Lancaster University. Retrieved from https://www.lancaster.ac.uk/fass/resources/sociology-online-papers/papers/jessop-g

Joshi, D. (2020, February 6). Welfare Automation in the Shadow of the Indian Constitution. Retrieved from Socio-Legal Review: https://www.sociolegalreview.com/post/welfare-automation-in-the-shadow-of-the-indian-constitution

*K. Puttaswamy v Union of India* (I) 10 SCC 1, 2017

Kahn, J. (2020, February 11). A.I. and tackling the risk of "digital redlining". Retrieved from Fortune: https://fortune.com/2020/02/11/a-i-fairness-eye-on-a-i/

Kaleidofin. (n.d.). About Us. Retrieved from Kaleidofin: https://kaleidofin.com/about-us/

Kannabiran, K. (2012). Tools of Justice: Non-Discrimination and the Indian Constitution. New York: Routledge.

Kar, S. (2018-a). Financializing Poverty: Labour and Risk in Indian Microfinance. Stanford University Press, 153.

Kar, S. (2018-b). Financializing Poverty: Labour and Risk in Indian Microfinance. Stanford University Press, 154.

Khaitan, N. (2019, October 25). New Act UAPA: Absolute Power to State. Retrieved from Frontline: https://frontline.thehindu.com/cover-story/article29618049.ece

Khaitan, T. (2009). Reading Swaraj into Article 15: A New Deal for the Minorities. NUJS Law Review.

Khanikar, S. (2018). State Violence and Legitimacy in India, 321.

Khilnani, S. (2004). The Idea of India. New Delhi: Penguin.

Kleinsteuber, H. J. (n.d.). Self-regulation, Co-regulation, State Regulation. Retrieved from https://www.osce.org/fom/13844?download=true

Kriebel, D., Tickner, J., Epstein, P., Lemons, J., Levins, R., Loechler, E. L. . . . Stot, M. (2001). The Precautionary Principle in Environmental Science. Environmental Health Perspectives, 871-876.

Kumar, A., Shukla, P., Sharan, A., & Mahindru, T. (2018). NationalStrategy-for-AI-Discussion-Paper. NITI Aaygo. Retrieved from https://niti.gov.in/writereaddata/files/document_ publication/NationalStrategy-for-AI-Discussion-Paper.pdf

Langenbucher, K. (2020). Responsible A.I. Credit Scoring – A Legal Framework. 25 Euro. L. Rev. 1.

Lending Kart (2020). Retrieved from Lending Kart: https://www.lendingkart.com/

*Lloyd Corp Ltd v Tanner*, 562, U.S. (1953)

Loan Frame (2020). Retrieved from Loan Frame: https://www.loanframe.com/

Maneka Gandhi v Union of India, SCR (2) 621 (1978)

Marda, V., & Narayan, S. (2020a). Data in New Delhi,s Predictive Policing System. Proceedings of ACM Conference on Fairness, Accountability, and Transparency. Barcelona, Spain, ACM, New York, NY, USA. USA. Retrieved from https://doi.org/10.1145/3351095.3372865

Marda, V., & Narayan, S. (2020b). Data in New Delhi's Predictive Policing System. Proceedings of ACM Conference on Fairness, Accountability, and Transparency, (p. 321). Barcelona, Spain. ACM, New York, NY, USA. USA. Retrieved from https://doi. org/10.1145/3351095.3372865

Marda, V., & Narayan, S. (2020c). Data in New Delhi's Predictive System. Proceedings of ACM Conference on Fairness, Accountability, and Transparency, (p. 322). Barcelona, Spain, ACM, New York, NY, USA. USA. Retrieved from https://doi.org/10.1145/3351095.3372865

Mittelstadt, B. (2019, May 20). AI Ethics – Too Principled to Fail? Nature Machine Intelligence. Retrieved from Nature Machine Intelligence: https://papers.ssrn.com/sol3/ papers.cfm?abstract_id=3391293

*Mullin v And'r, Union Territory of Delhi,* India, 2 S.C.R. 516, 518 (1981)

Nag, R. (2016, June 10). How Matrix Backed FinTech Startup Finomena is Disrupting the $8 Bn Youth Loan Market. Retrieved from Inc 42: **https://inc42.com/startups/finomena/**

NASSCOM. (2018). Agritech In India – Maxing India Farm Output. Retrieved from NASSCOM: **https://www.nasscom.in/knowledge-center/publications/agritech-india-%E2%80%93-maxing-india-farm-output**

Nayak, N. D. (2015, May 3). Agricultural sector needs technological intervention to face challenges. Retrieved from The Hindu: **https://www.thehindu.com/news/national/karnataka/agricultural-sector-needs-technological-intervention-to-face-challenges/article7166263.ece**

NITI Aayog. (2018). National Strategy fir Artificial Intelligence. 33-34. Retrieved from **http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf**

Oswald, M. (2018). Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power. SSRN.

Palmer, S. (2008, July–October). Public Functions and Private Services: A Gap in Human Rights Protection. International Journal of Constitutional Law, 6(3-4), 585-60.

*Partap Singh (Dr) v Director of Enforcement, Foreign Exchange Regulation Act*, AIR SC 989 (1985)

Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

Pasquale, F. A. (2017). Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society. SSRN, 78. Retrieved from **https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3002546**

Pearson, J. (2017). AI Could Resurrect a Racist Housing Policy. Retrieved from **https://www.vice.com/en_us/article/4x44dp/ai-could-resurrect-a-racist-housing-policy**

Pelaez, V. (2019). The Prison Industry in the United States: Big Business or a New Form of Slavery? Global Research. Retrieved from **https://www.globalresearch.ca/the-prison-industry-in-the-united-states-big-business-or-a-new-form-of-slavery/8289**

Pichai, S. (2018, June 7). AI at Google: Our Principles. Retrieved from Google: The Keyword: **https://www.blog.google/technology/ai/ai-principles/**

*Pischke v Litscher*, 178 F.3d 497, 500 (7th Cir. 1999)

Prince, A., & Schwarcz, D. (2020). Proxy Discrimination in the Age of Artificial Intelligence and Big Data. 105 Iowa Law Review 1257. Retrieved from https://ssrn.com/abstract=3347959

Randazzo, A. (2013). Can a Disruptive Fin-tech create a Mass Market for Savings and Investment in India? Retrieved from Kaleidofin: https://kaleidofin.com/kaleidofin-can-a-disruptive-fin-tech-create-a-mass-market-for-savings-and-investment-in-india

Ranger, C. (2018, November 13). Using machine learning to improve lending in the emerging markets. Retrieved from Harvard Business School - Technology and Operations Management: https://digital.hbs.edu/platform-rctom/submission/using-machine-learning-to-improve-lending-in-the-emerging-markets/

Rao, N. (2013). Three Concepts of Dignity in Constitutional Law. Notre Dame Law Review, 200.

Reddy, B. D. (2016, June 9). Microsoft, Icrisat develop new sowing app for farmers using AI and Azure cloud. Business Standard. Retrieved from https://www.business-standard.com/article/companies/microsoft-icrisat-develop-new-sowing-app-for-farmers-using-al-and-azure-cloud-116060900752_1.html

Rouvroy, A. (2013). The End(s) of Critique: Data Behaviourism versus Due Process. In M. Hildebrandt, & K. De Vries, Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology.

Scherer, M. U. (2016). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. Harvard Journal of Law & Technology, 29, 354, 357, 259.

Schulz, W. (2006). Final Report: Study on Co-regulation Measures in the Media Sector, Study for the European Commission. Directorate Information Society and Media . Retrieved from http://ec.europa.eu/avpolicy/docs/library/studies/coregul/final_rep_

Schulz, W., & Held, T. (2001). Regulated self-regulation as a form of modern government. Indiana University Press.

Scott, C. (2017a). The Regulatory State and Beyond. In Regulatory Theory, Foundations and Applications (p. 269). Australia: ANU Press.

Scott, C. (2017b). The Regulatory State and Beyond. In Regulatory Theory; Foundations and applications (pp. 269–270). ANU Press.

Sethia, A. (2015, March 21). "The BCCI Case on "Public Function" and Its Implications on Sports Governance. Retrieved from iconnectblog: http://www.iconnectblog.com/2015/03/bcci-case-on-public-function/

Sharma, S. (2018, July 9). How ISRO is helping in Uttar Pradesh Map and Predict Crime. Retrieved from Tech Circle: https://www.techcircle.in/2018/07/09/how-isro-is-helping-uttar-pradesh-police-map-and-predict-crime/

Sharma, V. (2017, September 23). Indian Police to be armed with Big Data Software to Predict Crime. Retrieved from The New Indian Express: https://www.newindianexpress.com/nation/2017/sep/23/indian-police-to-be-armed-with-big-data-software-to-predict-crime-1661708.html

*Siemen Engineering and Manufacturing Co. of India v Union of India*, AIR Sc 1785 (1976)

Singh, A., & Prasad, S. (2020). Artificial Intelligence in Digital Credit in India. Dvara Research. Retrieved from, https://www.dvara.com/blog/2020/04/13/artificial-intelligence-in-digital-credit-in-india/

Sinha, A., & Mathews, H. V. (2020). Use of algorithmic techniques for law enforcement: An analysis of scrutability for juridical purposes. 55(23). Retrieved from, https://www.epw.in/journal/2020/23/special-articles/use-algorithmic-techniques-law-enforcement.html

Smith, C. A. (2018). The Colour of Creditworthiness: Debt, Race, and Democracy in the 21st Century. Baltimore, Maryland: Johns Hopkins University. Retrieved from https://jscholarship.library.jhu.edu/bitstream/handle/1774.2/60992/FORSTER-SMITH-DISSERTATION-2018.pdf?sequence=1&isAllowed=y

*State of Punjab v Balbir Singh*, 3 SCC 299 (1994)

*Sundar and Ors v State of Chattisgarh,* 7 S.C.C, 547 para. 73 (2011)

Terry, N. (2019). Of Regulating Healthcare AI and Robots. Yale Journal of Law & Technology, 21, 18. Retrieved from https://yjolt.org/sites/default/files/21_yale_j.l._tech._special_issue_133.pdf

*Terry v Ohio*, 392 U.S. 1 (1968)

Travancore Rayons v Union of India, AIR SC 862 (1971)

UN ESCAP. (2019). Artificial Intelligence in the Delivery of Public Services. Retrieved from https://www.unescap.org/sites/default/files/publications/AI%20Report.pdf

*Zee Telefilms v Union of India*, AIR, SC 2677 (2005)

**Appendix: Examples of Regulatory Tools for AI**

## Accountability, Oversight, and Redress

- Clear, funded, and appropriate mechanisms for redress.
- Systematic and bottom-up impact assessment of potential harms to civil liberties and human rights.

- Detection, mitigation, and response mechanisms for possible errors as a result of initial training and self-learning.
- In-built audit mechanisms and possibility of verification by an independent third-party.

- Clearly articulated liability structures for situations that involve the use of an AI system.
- Mechanisms for consistent and regular evaluation and review of AI systems, including inclusive and bottom-up mechanisms for tracking impact.
- Communication of changes to AI systems resulting from monitoring and evaluation.
- Capacity-building and awareness of data-driven decision making in courts at national, regional, and district levels.
- Clear framework for working with the private sector, including enabling access to training data held by the private actor, opening up source code, and assigning clear modes of contractual liability.
- Certification schemes and trainings for end users.

## Equality, Dignity, and Non-discrimination

- Anti-discrimination standards in compliance with constitutional and international human rights laws.

- Diversity assessment for members of development/implementation team.
- Written standard operating procedures (SOPs) during curation of the data and training of the algorithm.

- Mechanism for incorporation of citizen voices and feedback throughout implementation.
- Framework for assessing disparate impact on specific vulnerable communities.

## Safety, Security, and Human Impact

- Impact assessment of all cyber threats to which the AI system could be vulnerable.
- Risk assessment towards identifying unintended consequences prior to development, including in unpredictable environments.
- Existing cyber security frameworks at a national level.

- Depending on the severity of impact, clear safety controls for a human to override the AI system or reject a prompt, recommendation, or decision by the AI.

- Regular security audits, patches etc.
- Framework for data breach notifications and bug bounty programs.

## Privacy and Data Protection

- Compliance with national and global protocols on data protection and governance, including consent principles, control over data use, and restriction of processing, right to erasure, and rectification.

- Clear regulatory frameworks for personal and non-personal data in existing data sets.
- Adoption of necessity, proportionality, and "least intrusive" standards to guide the design, development, and use of AI systems.
- Built-in mechanisms for notice and consent, with possibility to revoke.
- Ethical practices in collecting and accessing data for training purposes.

- Oversight mechanisms for collection, storage, processing, and use – particularly for real-time and long-term collection and use of data.
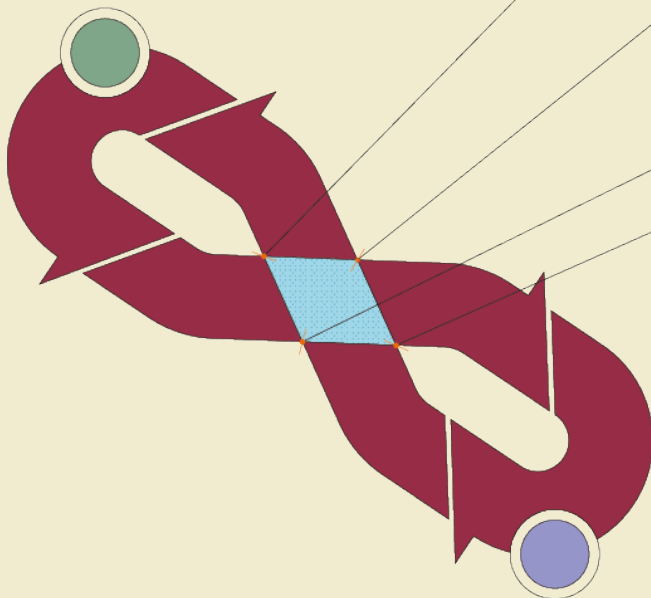
## Agency

- Comprehensive notice framework that accounts for passive and active data collection.
- Comprehensive transparency frameworks for data inputs, data training and curation, and use of decisions.
- Retrospective adequation.
- Opt-out options for individuals.
- Gradients of human-in-the-loop.
- Standards for accuracy.

# AI Technologies, Information Capacity, and Sustainable South World Trading

**Mark Findlay**
Centre for AI and Data Governance,
School of Law,
Singapore Management University

## Abstract

This paper represents a unique research methodology for testing the assumption that AI-assisted information technologies can empower vulnerable economies in trading negotiations. This is a social good outcome, enhanced when it also enables these economies to employ the technology for evaluating more sustainable domestic market protections. The paper is in two parts. The first presents the argument and its underpinning assumption that information asymmetries jeopardize vulnerable economies in trade negotiations and decisions about domestic sustainability. We seek to use AI-assisted information technologies to upend situations where power is the unfair discriminator in trade negotiations because of structural information deficits, and where the outcome of such deficits is the economic disadvantage of vulnerable stakeholders. The research question is the following: How is power dispersal in trade negotiations, and consequent market sustainability, to be achieved by greater information access within the boundaries of resource limitations and data exclusivity? The second section is a summary of the empirical work which pilots a more expansive engagement with trade negotiators and AI developers. The empirical project provides a roadmap for policymakers convinced of the value of the exercise to then adopt the model reflections arising out of the focus groups and translating these into a real-world experience. The research method we propose has three phases, designed to include a diverse set of stakeholders – a scoping exercise, a solution exercise, and a strategic policy exercise. The empirical achievement of this paper is the validation of the proposed methodology through a "shadowing" pilot method. It explains how the representative groups engaged their role plays, and summarizes general findings from the two focus groups conducted.

## Analytical Purpose

This paper represents a unique research methodology for testing the assumption that AI-assisted information technologies can empower vulnerable economies in trading negotiations. This is a social good outcome, enhanced when it also enables these economies to employ the technology for evaluating more sustainable domestic market protections.

The paper is in two parts: the initial discursive analysis presents the argument underpinning the assumption; the second section is a summary of the empirical work which pilots a more expansive engagement with trade negotiators and AI providers. This division allows a policy audience to concentrate on the justifications for the assumption, the challenges facing implementation, and the speculated consequences from its successful achievement. Researchers and evaluators will find interest in the details of the pilot methodology.

The paper demonstrates and tests our confidence in the methodology to positively establish the analytical assumptions regarding power dispersal and sustainable domestic market analysis. We advance speculative policy recommendations that can be drawn for the critical experience of the pilot methodology. The paper's commitment to empowerment through policy engagement and recipient ownership makes prescriptive policy inappropriate without a full application of the method in real market decision-making.

Consistent with the overarching project brief, we have identified a need and proposed an AI-assisted answer to that need at theoretical and policy levels. As such, a social deficit is established and a social good through AI is proposed, which is consistent with a major head of the ESCAP development goals. Recognizing resource limitations and time constraints, the empirical project in the second part provides a roadmap for policymakers convinced of the value of the exercise, to then adopt the model reflections arising out of the focus groups and translating these into a real-world experience.

In more detail, the policy and research assumption is that by employing AI-assisted information sourcing, sorting, and analyzing technologies to improve information access and evaluation underpinning economic decision-making, vulnerable economies can better determine sustainable domestic market policy against enhanced trade bargaining capacity. The availability of AI information-assistance technologies (and associated expertise/education)[1] will, it is argued, provide the material and understandings necessary (but currently absent or under-developed) for selecting contexts of domestic market protection to promote sustainability, and for more competently valuing trade bargaining positions in the case of transnational exchange markets.

At a more macro consideration of economic reliance, this policy decision-making enhancement will reduce the reliance on market surplus dumping from more powerful trading partners and its anti-subsistence consequences. As domestic market sustainability is more strategically prioritized, these vulnerable economies will better weather post growth, or de-growth global economic trends.

As for enhanced trading capacity, and specifically empowered trade bargaining positioning, AI information-assistance technologies for data access, automated data management, and analysis, it is argued, will offer social good outcomes to presently disempowered multi-stakeholder trading players who currently negotiate under information deficits and resultant weakened bargaining capacity. AI information-assistance technologies will strengthen bargaining power, which will increase trading revenue and make more achievable aspirations for "world peace through trade" (Dikowitz, 2014).

---

1. It is not the intention of the paper to specify these technologies. In fact, essential for our belief in recipient "ownership", any eventual policy applications should involve recipient economies in a dialogue with AI technical resource personnel and donor agencies, to determine the technologies best suited to need on a case-by-case basis.

## Background

The foundations of our thinking grow from the following propositions, which can be viewed as policy underpinnings:

1. General principles can be identified as governing successful trading bargains;[2]
2. Trade negotiations usually reflect the relative market power and positioning of participants;
3. Trading partners from more vulnerable economies may require external bargaining support if structural power asymmetries are to be dispersed in their favor;
4. A "free trade model"[3] has negative impacts in weaker economies being required to open up their markets and remove protections over domestic social production.[4] This trade liberalization has meant that domestic market subsistence and economic sustainability are diminished in favor of trading exploitation;
5. Weaker economies have been adversely affected by discriminatory trading arrangements and exclusionist trading alliances, particularly as their trade commodities are undervalued, and their attractiveness as preferred partners is equally so;
6. Automated data management[5], access to big data[6], and artificial intelligence technology capabilities[7], if affordably available to weaker trading economies, offer capacities to strengthen their positioning in certain trading arrangements;
7. A protectionist regression in domestic trade arrangements among major trading powers, and moves from multi-lateral to bi-lateral trading alliances, both designed to reduce individual trade deficits and to penalize offending trading partners, may offer opportunities for weaker trading economies to assert domestic social production and bi-lateral advantage. The reasoning behind this view is that domestic market liberalization North to South World, ignoring how vulnerable may be the target domestic resource market, leaves vulnerable economies even more exposed to trade discrimination when major global trading nations are reverting to selective and self-interested tariff protectionism;
8. The paradox between free trade open market liberalization, and intellectual property and data transfer protection, disadvantages weaker economies with lower levels of IP "ownership" and effective data transfer controls.

Taking these fundamentals as given[8], the first part of the paper builds the following argument:

- Employing bargaining theory, a typology of successful trade bargaining can be established and the significant factors, prioritized;
- Anticipating that information deficit regarding key aspects and dimensions of any particular trade bargain will further disadvantage weaker parties,[9] access to information and critically appreciating its analytical value will level the bargaining power asymmetries;

---

2. What is meant by "trade bargains" or "trade negotiations" here is specific trade deals rather than prevailing or permanent trade agreements and partnerships.

3. As a policy to eliminate discrimination against imports and exports, the free trading model has never fully been achieved globally. In such an ideal trading frame, buyers and sellers from different economies may voluntarily trade without a government applying tariffs, quotas, subsidies, or prohibitions on goods and services. Free trade is, therefore, proposed as the opposite of trade protectionism or economic isolationism. Instead of freedom and fairness, having attained comparative advantage in production, the hegemon is typically impaired by artificial trade barriers in its quest to penetrate the domestic economies of competing states. Thus, as a state rises from the core to hegemony, it will progressively favor lower tariffs and move towards a free trade doctrine for import receiving markets, while at the same time resorting to tariffs on imports where they are deemed to correct trade imbalances against their benefit. In de Oliver M. (1993) "The Hegemonic Cycle and Free Trade: the US and Mexico" Political Geography 2/5: 457-474.

4. Can social production at home be an adequate substitute for market production from producers abroad, particularly when it comes to high-tech commodities and services? The same could be asked about specialist natural resources which are the material life blood of high technology, and as such, trading priorities. We advance here that trade is necessary for balanced development, but trade deals need not crowd out domestic social production through the export dumping of subsidized or cheap replications of sustainable domestic social production.

5. This refers to the application of algorithmic technologies in cataloguing and mapping data at rest and in action, thereby lessening the prospect of "drowning in big data", https://erwin.com/blog/automated-data-management-stop-drowning-data/

6. The term "big data" has come to mean some form of "value-added" data application potentials. Simply, big data refers to extremely large datasets which may be analyzed computationally to reveal patterns, trends, and associations, particularly concerning human behavior and interactions. The size of these sets and their capacity to cross fertilize creates negative challenges to evaluating data sources and their progressive integrity.

7. The paper prefers the definition provided by Stuart Russell and Peter Norvig (2010) Artificial Intelligence: A Modern Approach (3rd edition), New Jersey: Prentice Hall; "the designing and building of intelligent agents that receive precepts from the environment and take actions that affect that environment". This approach connects with a key idea relevant to the present discussion, that AI is not the same as information – it is technology that helps us process information to take actions in the world.

8. It is possible for each of these assumptions to be empirically tested and contextually validated. However, for our initial purposes, they are designed to form the foundations of wider analytical projections.

9. Rather than talking about economies in terms of stages of development, this paper distinguishes participation in economic decision-making and trade bargaining in terms of the relative strength and weakness of participants. Vulnerability is the approach taken here as an empirical measure of relative market power, which can be corrected through more equal access to the information underpinning strategic economic decision-making.

- Understanding the dynamics of a global free-trading model, and its critique in the recent return to protectionism, projections could be offered regarding how weaker trading economies might be advantaged by interventions to improve their individual bargaining power, and at the same time strategically protecting their sustainable domestic social production;
- Information deficits regarding crucial trade bargain variables disadvantage parties[10] with reduced or restricted access to such information;
- Automated data management, access to big data using artificial intelligence technology, and enhanced analytical expertise/education can provide external assistance to disempowered trading parties when seeking to improve their bargaining status;
- Such information access capacity is made more viable through enhanced internet access;
- Aid and development agencies, international organizations, and private philanthropic entities can provide the financial backing to finance the necessary technology for trade information empowerment. Additionally, multi-stakeholder trading arrangements could fund AI information technology capacity to advance aspirations for "world peace through trade";
- Access to information alone will not rebalance trading power asymmetries. Along with more access, there is a need to invest in critical and resilient analytical capacity.

Each of the paper's policy underpinnings represent commitments to the greater trading sustainability of small and less powerful trading economies, in a global context where these economies can teach the North World much about sustainability in a post growth, or de-growth trading age. In addition, more encompassing policy eventualities directed to sustainability for vulnerable economies will be enriched by this research through the suggested potentials it offers to enhance informed decision-making about what domestic resources should be retained in domestic markets, and where these market can be opened up to trade without endangering the resilience of such economies.

# Part I

## The Analytical Challenge

Trade has become essential for the viability of today's exchange economies, big and small. Global trade that produces benefits for all is also seen as a positive aspect of global governance and peacemaking. Commodities traded will vary, largely depending on the demographics of the economy and its historical development. If we accept that "property is a fundamental social practice" and "ownership is indeterminate" (Humbach, 2017) then there needs to operate a sustainable frame for things traded between parties that want what property and ownership they claim, to work best for their complex social needs.

Unfortunately, as Joseph Stiglitz has observed at the forefront of free trade policy marketing operating from a beggar-thy-neighbor perspective to beggar-thyself (Stiglitz, 2002a), the "free trade" panacea did not realize universal benefits across the globe.

*International economic justice requires that the developed countries take action to open themselves up to fair trade and equitable relationships with developing countries without recourse to the bargaining table or attempts to extract concessions for doing so (Stiglitz, 2002b).*

Implicit in this recognition of requiring fair trade initiatives driven from the rich and powerful down to the poor and powerless, is pragmatic structural and process cautioning about unequal bargaining relationships. The cynic might say that fair trade is a non-sequitur. A good bargain benefits one to the detriment of the other. If this is the inevitability of trade, at a global level it explains the inequitable and destructive trajectories of contemporary global economic imperialism (Hardt & Negri, 2001). This paper does not proceed within any such inevitability. Nor does the paper ignore that the introduction of AI-assisted information technology can have the

---

10. Parties to economic decision-making and trade negotiations may be state actors, commercial agents, or multi-participant stakeholders.

unintended adverse consequences of increasing unfairness if the nature of trading biases based on wider hegemonic disempowerment is not appreciated. Laws against protectionism and promoting free trade North to South worlds often give "fairness" a low priority. Along with more access to information, we would encourage the development of legal regimes respectful of, and not simply exploiting, global economic disparity.

When reflecting the problems associated with transferring misunderstood or misconceived concepts of "fairness" into complex socio-technical systems, Xiang and Raji conclude that "fairness" is a mutual enterprise between AI-creators and legal policymakers:

> *If the goal is for machine learning models to operate effectively within human systems, they must be compatible with human laws. In order for ML researchers to produce impactful work and for the law to accurately reflect technical realities of algorithmic bias, these disparate communities must recognize each other as partners to collaborate with closely and allies to aid in building a shared understanding of algorithmic harms and the appropriate interventions, ensuring that they are compatible with real-world legal systems (Xiang & Raji, 2019).*

## New Global Economic Models

Sustainable world trade in an era of post growth or de-growth,[11] is facing challenges from the push for protectionism and isolationism against trade liberalization and the "wealth of nations". National self-sufficiency has incrementally been downgraded by free trade imperatives in favor of the internationalization of economic activities. Populist backlash would selectively reverse the forces of global economic engagement in preference for trading imperatives governed by domestic surplus and offshore relative disempowerment.

The potential downsides of free trade are said to be mitigated by:

- Allowing for innovation and structural change;
- Increasing employability and enabling life-long learning; and
- Redistributing globalization gains more-equally in domestic economies through taxation (Reichel, 2018).

Debate these eventualities if you will, but their achievements are no doubt dependent on which side of the globalization engine one sits – is it for prosperity and peace, or alternatively, for intra-country wealth through production chains skewed to stronger economic bargainers?

The political and economic reality of current trade agendas is that vulnerable economies will be negatively impacted via protectionist policies enforced by major trading nations, in different ways but to similarly disabling extents as they were when forced to expose their own markets to the unbalanced influence of North World free trade expansionism. The inequalities of free trade and selective protectionism, operating on profound imbalances in trade capacity, represent the context for policy reform advocated in the remainder of the paper.

## Specifically, the policy reform advocated in this analysis involves:

- Recognizing that sustainable global economies will not be advanced by a heavy regression to selective protectionism or a blind adherence to discriminatory and unbalanced trade liberalization.
- Appreciating that free trade can continue as a dimension of positive global engagement where free trade agreements allow for domestic social production and thereby advance the aspiration for world peace through trade.

---

11. These are several definitions of de-growth which largely focus on economic policy which concentrates less on economic stimulus than sustainable social welfare. For this paper, the concept also incorporates "post-growth" – economic economic inevitabilities which see growth slowing or flattening irrespective of political and market intervention. See Azam G. (2017) "Growth to De-Growth; a brief history" https://www.localfutures.org/growth-degrowth-brief-history/. "[De-growth] challenges both capitalism and socialism, and the political left and right. It questions any civilization that conceives freedom and emancipation as something achieved by tearing oneself away from and dominating nature, and that sacrifices individual and collective autonomy on the altar of unlimited production and the consumption of material wealth. Capitalism has brought further ills such as the expropriation of livelihoods, the submission of labor to the capitalist order and the commodification of nature, (for the South World in particular). This project to establish rational control over the world, humanity and nature is now collapsing."

- Realizing that the current financial sustainability of vulnerable South World economies, despite those being economies more likely to adjust successfully to post-growth or de-growth regimes,[12] will be enhanced if their bargaining power in trading arrangements, and their capacity to discriminate between what should be traded and what should remain a domestic resource, is empowered through greater information access and analysis.[13]

The next section looks at a model of bargaining dynamics. In particular, it identifies the importance of access to information for empowering bargain participants.

## Bargaining Theory[14]

What factors determine the outcomes of specific trade negotiations? What are the sources of bargaining power? What strategies can help in improving a party's bargaining power?

Trade bargains can be epitomized as at least two parties engaging for the purpose of some beneficial outcome (which might or might not be mutual) but who have conflicting interests over terms. These common interests are in cooperating for trade; the conflict lies in how to cooperate.

Taking a more contextual approach, understanding the dynamics of bargaining from the perspective of disadvantaged parties in particular, provides an opportunity to appreciate market dynamics and relationships (internal to the bargain) as well as the influence of political and economic policies' repositioning transactions (external). Interrogating the essential features of the bargain requires more than disentangling reasons for agreement or disagreement. A power analysis is at the core of bargaining theory, governing the imperatives for gaining the best benefit, and often at the cost of fairness or other more universal normative considerations.

Practically, issues of efficiency and distribution are important. Efficiency is at risk if the agreement fails or can only be reached after costly compromise and delay. Distribution relates to how gains emerge from co-operation between the two parties. To these issues identified by Muthoo, we would add sustainability. It is rare that trade relationships are "one-off's". They usually lead on to the establishment of enduring market connections, or they have ramifications for the parties involved, which stretch beyond the commercial terms of the deal.

What are the determinants of the bargaining outcome?

## A. Impatience, or the pressures of time

Each player values time. The preference is to agree to the price today rather than tomorrow. The value given to time will be subjective and relative. In particular, it may be as disproportionate and incremental as it is exaggerated by other external cost pressures. Weaker players may have less time to bargain or stronger players may exert the pressures of time if the rapid conclusion of the bargain is essential for other bargains to follow.

Apparent impatience can lead to a weakened bargaining posture or a breakdown of other rational communication essentials. In order to avoid the exposure of impatience, bargaining theory suggests that the vulnerable party should decrease their haggling costs and/or increase the haggling costs of the other party. *One way of achieving such a differential is for the otherwise impatient party to possess and understand the richest range of information and data that constructs (or constricts) the other party's bargaining context.*

Because the wealth and power differentials between trading parties are structural (and often not temporal or spatial), a basic principle of bargaining theory is that economies are unlikely to converge in wealth and income solely through international trading policy.

---

12. Some say that developing economies need the benefits of growth before adopting a largely North World economic countermovement like de-growth. There is an alternative argument that the conditions required for rethinking the place of the economy within the social, and prioritizing social rather than material goods, are more apparent and resilient in less modernized and less materially dependent societies. The debate is usefully discussed in Lang M. (2017) "Degrowth: Unsuited for the Global South?" *Alternautas*. http://www.alternautas.net/blog/2017/7/17/degrowth-unsuited-for-the-global-south. In any case, we are not requiring de-growth, but rather post-growth approaches to sustainability that accept growth as a priority for the South World but in the context that economic growth is repositioning as a global economic agenda.

13. In advancing this thesis, we are mindful that information access alone will not empower market stake-holding. The quality of that information (i.e., its relevance, immediacy, and analytical transparency) all depend on more than technological facilitation. The factors on which information empowerment relies are contextually important when evaluating the significance and sustainability of technological facilitation.

14. The following summary draws heavily on Muthoo, A. (2000) "A Non-technical Introduction to Bargaining Theory" *World Economics* 1/2: 145-166

*Features integral to bargaining dynamics such as information deficit, we argue, have greater potential to counterbalance prevailing structural inequalities that determine patience to let negotiations run their natural course.*

## B. Risk of breakdown

If while bargaining, the players perceive that the negotiation might break down into disagreement because of some exogenous and uncontrollable factors, then bargaining dynamics will alter. Risk of breakdown can be raised through a range of variables from human incompatibility, to the intervention of third parties.

This risk perception is where strategies to increase risk aversion are important. *Information available to parties concerning the nature of the risk and its impact on the other side becomes important if a weaker party wants to shield through risk aversion.*

## C. Outside options

Here, the principle is that a party's bargaining power will be increased if their outside option is sufficiently attractive – that is where alternative trading/ bargaining arrangements may parallel the first instance bargaining. Weaker parties are often devoid of any other option, outside or otherwise, or because of not fully understanding the values and variables at play in their bargain, feel trapped within a trade that is anything but to their advantage. *The outside option principle is directly impacted by the amount of information either or both parties have about the bargain in play and the outside option relative to the first instance bargain. The valuation of an outside option will depend not only on the conditions and characteristics of that option, but as much or more on its consequences for the bargain in play.*

## D. Parties' relationships

There is much in bargaining theory which concerns the significance of connections between the parties in contexts outside the bargain in hand. These externalities (such as cultural familiarity and political bonds) may impress so deeply into every other condition of the bargain, that negotiations cannot break free from responsibilities and obligations inherent within any such prevailing relationship.

*Again, information imbalance, or data access restrictions built into such extraneous relationships will further exacerbate the information deficit retarding knowledgeable participation in the eventual agreement struck.*

## E. Parties' interests and preferencing

Individuals and organizations seeking to influence economic decisions or to achieve success in a trade bargain, approach the enterprise with pre-formed preferences and exhibiting internalized interests. The decisions or bargains with which the result will be colored by such preferences and interests in the same way that any market choice is in part the product of preference gratification, interest, containment, or satisfaction. Pound (Grossman, 1935) would see the contest over interests as settling on individual claims, demands, or desires. How any of these features have a preference through a bargain or decision will reveal the relative power exercised by individual stakeholders, and by dominating any conflict over interests, the way power differential may be increased.

In trade negotiations, the interests of stakeholders will range well beyond the remit of what is to be bargained or decided. *Therefore, if the influence of pre-existing preferences and interests is going to weigh significantly on the negotiation or decision-making dynamics, then the more each party has detailed and informed knowledge about these preferences and interests, the less likely they will distort outcomes in ways which could not be planned for or at least anticipated by negotiating parties on both sides.*

## F. Commitment tactics

In many bargaining situations, the players often take actions prior to/or during the negotiation process which partially commit them to some strategically chosen bargaining positions. If these commitments are partial in that they are revocable, depending on how far down the line of negotiation they have been struck, this may progress the appearance of intractability and therefore costs associated with their revocation. Many of these commitments may have been orchestrated in order to increase the "bluff" (e.g., the limitations on a party to negotiate freely beyond the terms of another commitment). *The power of bluff is always dependent on contrary information or any suspension of disbelief in the bluff.*

## G. Asymmetric information

It might be accepted bargaining practice that one party will always know something the other does not. How such an information disparity should be valued is relative to the significance of the information for the vital terms of agreement (or disagreement). Information asymmetries affect both the values and pricing on offer as conditions of a deal, as well as when the agreement might be concluded for the maximum mutual benefit.

In general, an absence of complete information will lead to inefficient bargaining outcomes, even for those who benefit from an information surplus. The logic behind this view rests on an acceptance that the more information available to both parties, the earlier synergies will be established and bargains struck.

*The message is that treating information in some exclusionist or proprietorial manner may produce a short-term bargain benefit for the information owners (renters and possessors), but at the risk of an unsustainable trading market vulnerable to misrepresentation, exploitation, corruption, and the retarding on any natural propensity for market competition.*

*Therefore, policies to defeat information asymmetries in trading arrangements, we argue, offer empowerment potentials for weaker players, and on the strength of power dispersal through information access and sharing, more sustainable trading markets ongoing.*

In seeking power dispersal via information access and analysis, this paper is not requiring some egalitarian levelling of market engagement. As Rawls argued, social inequality will not always be the product of power abuse or discrimination (Grcic, 2007). What we are seeking to attack are those situations where power is the unfair discriminator because of structural information deficits, and economic abuse of vulnerable stakeholders is the outcome.

*From this review of bargaining dynamics, the essential research question emerges: How is power dispersal in trade negotiations, and consequent market sustainability, to be achieved by greater information access within the boundaries of resource limitations and data exclusivity?*

## Access through AI, Automated Data Management, and Big Data – Some Critical Considerations

As suggested in the brief reflection on fairness (above), it is necessary to preface any consideration of the relationship between improved data access, and improved bargaining power in trading arrangements, with the caution that more data and better automated data management courtesy of AI technologies will not automatically empower weaker trading partners. In fact, increased technological capacity to access data, unconnected with significant advances in data appreciation and contextualization may simply further fog the understanding of smaller stakeholders and exacerbate bargaining disempowerment.[15]

In addition, bargaining tactics may prefer privacy when information is applied, sought, withheld, or

---

15. The focus group discussions in the Methodology section enunciate this concern.

exchanged. The bargaining attitude that bargaining power is lessened if information is mutualized has to be addressed with the argument that for market sustainability, and not just a single bargain advantage, fairer information access will make for more robust economic engagement. Once again, we return to the externalities of economic fairness.

How market stakeholders accommodate and benefit from information abundance is at the heart of any policy derivatives designed to improve trading balance in a hegemonic global trading model, intensified in its potential to discriminate as a consequence of selective and politicized protectionism. A feature of the methodology to follow is the potential to better understand how information needs to be met with enhanced information access to address specific bargaining decisions.

An important consideration, which informs the policy projection for trade empowerment, is its timeliness. With the major trading partners at war over tariffs, trade imbalances, protectionism, and perversely, secrecy when it comes to tech transfer and IP, the conditions may be right for smaller trading economies to rebalance their domestic sustainability without the backlash of free trade essentialism.[16] From that stance, an informed and economic evaluation of what remains open for trading will provide a more stable platform for trade bargaining.

Access to information, complemented by increased analytical capacity, will enable more nuanced distinctions between protection for domestic sustainability and competitive positioning in regional and international trading. Yet, strategic analytical capacity does not simply depend on more devices and bigger technologies. In fact, the savvier information-users are navigating away from an over-reliance on devices and are becoming aware about how algorithms affect their lives. In any case, even those market players who have less information are relying

more on algorithms to guide their decisions, whether they realize it or not. In the current technologized world environment, it is axiomatic that new digital literacy is not about more skillfully using a computer or being on the Internet on call, but understanding and evaluating the consequences of an always-plugged-in lifestyle for every aspect of social and economic engagement. In societies and cultures that still place social relations much above digital connections, the introduction of AI capacity is never, as we see it, meant to diminish or downplay the dominant role of human agency.

Over two thirds of the worlds' population either live outside or can only partially participate in the digital age. Digital access and digital literacy are now recognized as fundamental human rights. However, when it comes to fair trading practice, a level playing field in terms of information engagement is not only a long way off, but some might argue is a misunderstanding of bargaining behavior and advantage (UNCTAD, 2019).

In 2014, the UN General Assembly adopted resolution 69/204 "Information and Communication Technologies for Development". Most relevant for this paper is the reference to:

> "…*information and communications technologies have the potential to provide new solutions to development challenges, particularly in the context of globalization, and can foster sustained, inclusive and equitable economic growth and sustainable development, competitiveness, access to information and knowledge, poverty eradication and social inclusion that will help to expedite the integration of all countries, especially developing countries, in particular the least developed countries, into the global economy" (UNCTAD, 2015);*

This paper is not solely concerned about a "digital divide" between those who have access to computers and the Internet and those who do not. As digital devices proliferate, the divide is not just about access

---

16. As noted earlier, there has been much political hypocrisy surrounding the "freedom" of free trade, and as such, a re-balancing of domestic sustainability and regional/international competitiveness will not necessarily require a wholesale rejection of more open cross border commercial engagement.

17. The mirror image of this divide is the incapacity of algorithm designers to appreciate the complexity and sometimes intentional ubiquity in the social circumstances and human decisions to which they are applied.

or available technologies. How individuals and organizations deal with information overload and the plethora of algorithmic decisions that permeate every aspect of their lives is an even more relevant discriminator when turning a power analysis to the global trade divide (Susaria, 2019). The new digital divide is wedged over understanding how algorithms can and should guide decision-making.[17]

*The "empowerment through data access and analysis" model that is advocated here depends on the availability of technological facilitation in identifying relevant data, determining its legitimacy and fitness-for-purpose, alongside enhanced analytical capacity and an upgraded appreciation of how AI as information technology can enhance essential economic and trade decision-making.*[18] Along with this external impetus for empowerment in decision-making is a concurrent challenge for information users in vulnerable economies to more clearly determine who decides what technologies should be preferred and whether such technologies offer decision-making options that are fair/legitimate/fit-for-purpose.

Syncing AI potentials with the information needed for domestic and trans-national trade bargaining and economic decision-making, is not singularly a question of sourcing and supplying technological capacity presently unavailable to weaker market stakeholders. Along with improved access and analytical technologies, there is a need to target the utility of such technologies and the information they produce to domestic economic sustainability (through trade protection) and increased trading profitability (through sharper trans-national bargaining).

In identifying the necessity for a more level playing field over data access and analysis in trade negotiations, this paper is not traversing discussions of "data trade" and its regulation, nor are we focusing on data driven economies.[19] The policy product of the research to follow is also not seeking to challenge even the most discriminatory IP and data protection regimes, though such challenges might successfully advance market sustainability in an era of access revolution (Findlay, 2017). Rather, the purpose of the research method to follow is to scope the type of information necessary for successful domestic market discrimination and trade negotiations, and the manner in which the provision of access and analytics technology (via AI potentials) can enhance the decision-making benefits which sustainable domestic market analysis and invigorated trading negotiations offer for empowering and assisting vulnerable economies at a time of world trade transition.[20]

## Bargaining-empowerment Through Technologized Information Access and Analysis

Bargaining-empowerment though information access may occur in several ways. Recognizing there is a difference between:

1. access to information helping an individual actor to bargain better, and
2. access to information assisting this actor to locate other stake-holder participants, and together they bargain better (because they share information and they act as a more influential bargaining unit);
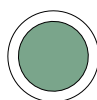
18. In identifying this decision-making "space", we recognize the importance of determining how to increase domestic market sustainability, while at the same time evaluating what should be traded beyond the domestic market and at what value.

19. For an interesting discussion of these two themes and their intersection, see Ciuriak D. (2018) "Digital Trade; Is data treaty-ready" *CIGI Papers No.162*
https://www.cigionline.org/sites/default/files/documents/Paper%20no.162web.pdf

20. In talking of trading arrangements in terms of state-to-state dialogue, we are, for the purposes of this research, simplifying the trading demographics wherein private sector players may be as significant or more so when vulnerable stakeholders in trade negotiations expose their domestic markets and resources to the interests of external multi-national traders. This paper was settled prior to the impact of the COVID-19 pandemic on global economic relations and as such cannot take these influences into account for this analysis.

21. We recognize that these quality-control problems are exacerbated the bigger and more interconnected are the datasets.

Once information has been identified, its sources need to be understood and the prudential pathways through which it has passed if relevance and reliability are to be measured.[21] The quality of information matters in terms of its decision-making value, and information offers diminishing decision-making returns as that quality is less open to testing and verification. A small amount of high-quality information is likely more useful than an abundance of low-quality information. In that regard, access and analysis must be accompanied by easy methods for data evaluation against simple matrices. An example of the variables to be considered would be (where visible) completeness, timeliness, uniqueness, accuracy, validity, and consistency (IT Pro team, 2020).

Next comes the issue of information over-load. Unleashing masses of information, high quality or not, will swamp vulnerable users without the capacity to process it. The other side of this problem is where AI and data analysis technologies can respond for social good.

Is this assertion confirmed by the literature? The studies associated with improved bargaining power as a consequence of greater information analysis are heavily concentrated on labor mobilization.[22] Analogies are usefully drawn from this literature insofar as it has a distinct interest in negotiations, bargaining, and decision-making modelling.

It is not novel to suggest that AI technologies can enable better trading outcomes for vulnerable economies. The United Nations Conference on Trade and Development (UNCTAD) recently introduced a new AI tool to speed up trading negotiations by simplifying complexity. As part of the Intelligent Tech and Trading Initiative[23], UNCTAD and the International Chamber of Commerce have produced a prototype of what they call the Cognitive Trade Advisor.

*"Developing countries and least developed countries have limited resources to prepare for trade negotiations,"* said Pamela Coke-Hamilton, Director of International Trade and Commodities of UNCTAD.

*"The amount of information that negotiators and their teams need to process is proliferating, and often they need the information on a timely and rapid basis,"* she said. The Cognitive Trade Advisor uses an understanding of natural language to provide cognitive solutions to improve the way delegates prepare for and carry out their negotiations.

*"The texts of the agreements are getting longer and longer,"* Ms Coke-Hamilton said. *"In the 1950s, an average trade agreement was around 5,000 words long. In the current decade, this has increased to more than 50,000 words. Dealing with such amounts of information takes a lot of time (UNTCAD, 2018)."*

Interesting as this development might be, our policy frame has a more restricted but no less impactful intention. As mentioned earlier, we are not touching on preferential trading arrangements, or the understanding of their complex documentation.[24] Instead, our remit is more contained, and as such, attainable without new technologies. The direction of the policy to follow is the employment of presently available AI technologies for accessing and analyzing information that can better position vulnerable negotiators by reducing crucial information deficits. The UNCTAD initiative is to develop new AI tools in order to make the attainment of Sustainable Development Goals more likely in under-developed regions. This paper shares the desire to see AI supporting progress to these goals by reducing negotiating inequalities. On the way to achieving this aim, the poverty in AI experience with currently available technologies in vulnerable markets and societies will hamper developments towards these goals even before new, affordable, user-friendly, and sustainable technologies are more readily available.

---

22. An example is https://turkopticon.ucsd.edu/
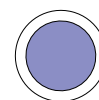
23. Information retrieved from https://itti-global.org

24. For example, see Alschner, W., Seiermann J., & Skougarevskiy, D. (2017) "Text-as-data analysis of preferential trade agreements: Mapping the PTA landscape" UNCTAD Research Paper No. 5. https://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=1838

Employing AI-assisted technologies for information access may or may not be in itself a neutral endeavor. In advocating this progress, there needs to be sensitivity to political and cultural parameters in offering AI technologies to analyze and prioritize economic and trading decision-making. Many post-colonial vulnerable economies do not respond well to top-down capacity building from the North World, especially when North/South disempowerment is identified in these economies as the root cause of their trade problems in the first place.

It is not the intention of this paper to provide a pre-packaged menu of preferred technological options to enhance access and analysis. As the methodology section to follow sets out, "ownership" of this selection should be offered through a scoping exercise which identifies context-specific needs and solutions. Ultimately, the preferred technology should be seen by the potential user as at base beneficial and manageable within the specific dynamics of their decision-making and bargaining ecology.

In seeking to identify the types of AI-assisted information technology that would best support vulnerable economies in domestic resource economic decision-making and trade negotiations, the following factors are important selection criteria and *determine how the policy suggestions in this analysis should be implemented:*

- *The technology needs to be affordable.* Even if its purchase is subsidized there are running and maintenance costs which will fall to the user and as such, these need at least to be defrayed by cost-savings through improved decision-making outcomes and bargain positioning.
- *It must be user-friendly* and explicable so that institutional, cultural, or administrative resistance to new technologies, or suspicions about the hidden agendas they might translate from donors, can be overcome.
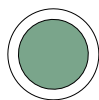
- *The technology must be robust and resilient.* The anticipated user population will not be sufficiently resourced with sophisticated tech support to manage frequent and constant hardware and software upgrading.
- *It should be capable of timely employment* in the various vital stages of decision-making and bargaining.
- It must have *rapid analytical capacities*.
- Its operational language must be i*n sync with the language of the bargainers and decision-makers.*
- On the basis of the information it accesses and analyses, *it should provide cognitive solutions from which the participants can draw informed choices.*

On the nature of information absent for access by policymakers looking at trade and domestic resource market sustainability from the perspective of vulnerable emerging economies, the imperial influence of platform distributors over raw data is an important reflection in the empowerment equation.

The commercialization and monetarized analysis of raw data through the big platforms presents a significant challenge when approaching the issue of more open access as an empowerment policy (UNCTAD , 2019). Accepting that there will always be sensitive metadata driving information technologies and linking through even simple keyword searching to an array of mediations over raw data for commercial purposes. The present project cannot neutralize this phenomenon, but it can flag it as a further level of potential disempowerment and seek transparency and explainability of data sourcing and technological translation in a language that the end user can appreciate and take into account when relying on information.

It is with this caution in mind that the project methodology is advanced.

# Part II

## Methodology

The project's methodology involves a pilot stage, the results of which are summarized in the conclusion of this section. Having satisfied ourselves that the focus group methodology is appropriate to test the analytical underpinnings, the project-proper methodology is described for later implementation.

The methodology has two clear underpinnings. First, to adopt a top-down approach to empowerment, with stakeholders already distrustful of the motivations which may underlie the actions of parties who in the past have been seen as complicit in the disempowerment reality, would endanger the sustainability of the support provided. Aligned with this is the second concern that both the research and the policy outcomes it supports should form stages in the empowerment process.

Therefore, the initial context for designing the first component of a research plan is to appreciate the nature of decision-making vulnerability that trading policy will need to address, and sustainability evaluations will need to constantly be monitored. Vulnerability is not to be viewed only in terms of power imbalance, or to substitute for terms such as "weakness", "disadvantage", and "discrimination" (Fineman, 2019).

Applying this individualist conceptualization of vulnerability to economies, markets, and societies, we can imagine a research method that appreciates the forces which create and maintain vulnerability, and provides a voice to the disempowered that resultant policy is designed to enable. In particular, and working from our earlier review of bargaining theory, the research should test whether decision-makers from

vulnerable economies realize information deficit in terms of decision-making need, can articulate the sources and substance of information that would be useful to them, and from there speculate on how such information wants to be analyzed, validated, and sustained. Once this "needs analysis" has been trialed, it then becomes the task for information technologists, with an understanding of information disadvantage and its decision-making context, to suggest AI-assisted information options that could empower sustainable decision-making.

The research design in its post-pilot phase has three phases:

### Participant Focus Groups

In the format of a facilitated focus group, a series of hypotheticals designed to provoke situations of vulnerability in economic decision-making and trade bargaining will be put to a meeting of negotiators and policymakers who have experienced disempowerment through information deficit. Recognizing the risk of "digital imperialism" in designing research experience from an external AI focused context, these hypotheticals will have been previously discussed, critiqued, and settled by a small working party drawn from scholars, negotiators, and policy people with a familiarity of South World economic disempowerment. In particular, the advice on drafting the hypotheticals will be taken in the first instance from experts in mediation and negotiation with hands-on experience of South World decision-making styles. Added to this will be the impressions from policymakers and negotiators working in South World trade and economic environments.
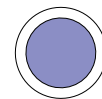
## a) Policymakers Focus Group, Scoping Exercise

Participants in this focus group will be drawn from five nominated economies with currently unsustainable domestic resources, limited trading advantage, and who it might be argued, have suffered as a consequence of North/South World market liberalization.[25] The output from this focus group would be a clear understanding of the information needs of participants, the situations in which the absence of specific information on which to rest decisions or make bargains is deemed to disempower, suggestions concerning what information access needs to be prioritized, and the types of cognitive options that participants would think helpful and why.

## b) Expert Focus Group, Solution Exercise

Armed with the information disadvantages identified in the first focus group, experts in the field of AI technology, development capacity building, trade negotiation and economic decision-making, mediation, and multi-participant stakeholder policy work would be charged to apply particular AI technologies to the problems represented in the first set of hypotheticals. In addition, members of focus group 2 will have access to a structured transcript of the discussions emerging out of focus group 1. Using the same hypotheticals across both focus groups will offer some qualitative consistency and comparability. Participants in the first focus group could be invited to attend and observe these discussions. The output from this focus group would be the preparation of a set of AI technology options nominated against the particular information deficits identified by the first focus groups. In preparing and costing these options, participants would be asked to reflect on the list of selection criteria that is described above.

## c) Implementation Focus Group, Strategic Policy Exercise

The final focus group would involve academic experts in vulnerability and social justice, as well as negotiation/mediation, policy regulation, and social development, similar to those drawn together to formulate and test the hypotheticals. The Emory/Leeds Vulnerability Initiative and scholars with interests in law and development, negotiation and mediation, and information systems would facilitate a policy forum designed to produce a workable policy agenda for information empowerment and market sustainability in the five nominated vulnerable economies. Additionally, experts from global information and communication organizations with abilities to fund a pilot scheme, representatives from ESCAP with responsibilities for promoting the UN Sustainable Development Goals, and interested participants from the previous two focus groups, would add to the policymaking dynamics. The policy yield from this workshop would be to roll out a pilot program that would enable an empirical evaluation of the impact of AI technology capacity building on the achievement of better trade bargaining benefits, and sustainable economic decisions regarding the safeguarding of domestic resources in vulnerable economies.

## "Shadowing" Pilot Focus Group Method-validation Exercise[26]

"Shadowing" is a style of simulation, where the survey population is brought together (usually at a pilot stage) to represent the intended actual survey population for the purposes of testing whether the research methodology is promising and potentially reliable. Shadow methodology is where the survey population is asked to assume the roles and responsibilities of an actual population, and where possible, to follow the progress of that population as it performs a particular decision task. This method has a history in jury research, in the US.

For the purposes of piloting, a combination simulation/shadow methodology was applied through two focus groups. The first identified information deficiencies in trade bargaining and domestic resource sustainability among trade and development policy personnel. The

---

25. APRU member institutions and their affiliates will be helpful in identifying and facilitating participants.

26. Due to pressures of time and limited resources, the pilot was not able to target policy makers in vulnerable economies (focus group 1). However, it was possible to engage with young AI technical experts in focus group 2. The implementation focus group 3 was not necessary at the pilot stage.

second group, with the benefit of such identification, then offered AI-assisted technology options. Hypotheticals enabled information and decision-making simulation to be experienced and monitored, and assembling a group of participants who were instructed "in-character" provided the "shadow" survey population with capacity to identify information disadvantage prompted by the hypotheticals.

For both groups, a small number of participants with similar social demographics[27] were drawn together.[28] The first group was asked to take on the character of a trade negotiator, or a trade and development policy officer in a nominated emerging economy.[29] Along with the identified role and jurisdiction, each participant was assigned a particular strategic concern in performing their function. That concern was connected back to limitations in the information base, or information deficits effecting the potential of each player to make knowledgeable trade policy or bargaining decisions, and determinations about domestic resource market sustainability.

From that perspective, and prior to focus group 1, each participant was encouraged to research their role with limited direction from the focus group administrators. The reason that this research stage is unstructured relates to the expectation that members of the actual population (trade policymakers, trade bargainers, and domestic resource decision-makers) will possess different levels of knowledge and experience depending on personal, professional, and information-centered variables, as well as differing degrees of self-reflection.[30] The only direction given for this independent, individual research phase was the necessity to focus on what information sources, technologies, and analytical capacities exist in the nominated jurisdiction. Even the actual population would be differentially challenged to know and identify what information is missing and what is needed due no doubt to varied personal experience and confronting variables (structural and functional) that are sometimes difficult to enunciate.

The second focus group was drawn from participants with special skills in AI-assisted information technology options. This group, while not specifically researching the information disadvantages existing in the 5 vulnerable economies, were required to reflect on these in a general sense and were assisted by the transcript from the first focus group, as a discussion and reflection resource.

The hypotheticals designed for group 1 to elicit information deficit and information need, contextualized knowledgeable decision-making and empowered bargaining/resource-retention determinations in three specific directions: natural resources trading, consortium-sponsored foreign direct investment, cash cropping diversification and regional security (see Appendix 1).[31] With the identification of information/analysis need, focus group 2 were asked to suggest and design practical options from available and affordable AI-assisted technologies directed to trade bargaining, and trade/domestic resource balance. Sustainability for these options is a priority.

Simulation/shadow survey populations are a compromise at the pilot stage but, with the participants applying sufficient dedication and immersion to their role-play, the discussions unfolded as a useful test-pad for whether this method should be applied in the more resource-demanding environment of actual survey populations. In particular, we wanted to explore whether participants can see the issues with what should be available to them, what they do not know, what is hidden, what-leads-to what in any information chain, and once more information is available, how it can empower the decision-making/analytical challenge. It proved possible to elicit responses along these lines in the context of the hypotheticals (see General Findings). It also emerged possible from group 1 to group 2 that information deficit, once identified, was followed by technological enhancement, which will lead to more empowered bargaining/decision-making capacity and outcomes.

---

27. A group of young, tertiary educated men and women with varied knowledge of the essential population experience, but briefed to take on a character within a defined context.

28. Once the participants for focus groups 1 and 2 were identified, they were separately briefed as to the purpose of the shadow simulation and were assigned characters and tasks to research and adopt.

29. The economies selected were Papua New Guinea, the Philippines, Vietnam, Cambodia, and Myanmar.

30. The need for self-reflection is a central tension in the exercise of any focus group method.

31. The testing of hypothetical utility would be another feature of the focus group experience and ownership.

## Focus Group 1 – General Findings

Starting out with the MNC/natural resource trading scenario, the initial information need centered on sufficient knowledge about the bargaining partner and the possibility of developing a relationship of trust. In addition to what might be found on the commercial public record, it was suggested to use already existing public and private sector trading networks, and explore previous case-study instances of the operation of the MNC in the region with similar trading conditions. Reservations were expressed about asking the MNC directly, based on different interpretations of power imbalance.

Next, it was deemed necessary to identify major decision sites and bargaining points in the commercial supply chain if the deal progressed. It was noted that some information along the chain might be protected as commercial knowledge. Mention was made about information access costs (material and representational) in contacting third parties and seeking commercial data. Would there be available historical aggregated data on harvesting, processing, marketing, and consumption and where and how could it be accessed? There seemed limited possibilities across other government and private sector agencies in each economy as the natural resource in question was yet to be commercially exploited. International organizations may have relevant data, but because each economy was not already linked to the international standardization networks for this resource, this information might not be easy to access.

Recognizing that this bargain had to be considered against competitive offers (or even exploitation by the state itself) how could other potential investors be approached without damaging the confidence of the deal on the table? What information would be necessary to identify markets for the natural resource, possible market prices, and features of alternative deals that should be anticipated?

Much of this information could come from the MNC itself, but commercial confidentiality may limit this as a source. In any case, information from a trade bargainer would require third party validation. How might this be achieved?

Several participants wanted to ensure that any such trade deal should be the first stage in a commercially sustainable arrangement. Aligned with this concern was interest in the sustainability of the natural resource, and the impacts on a pre-existing subsistence economy relying on the natural resource. Without any detailed natural resource surveys or environmental impact evaluation capacity, what were external analysis options to fill these information deficits?

Assuming that information regarding the MNC, the supply chain, and resource sustainability are available, the group discussed other information needs that impacted on relative bargaining power. It appeared that previous experience in natural resource trading was an important viable. Furthermore, general prevailing trade policy impediments such as nationalized industrial development, institutional corruption, exchange rate interference, and weak trade positioning against major trading economies were identified.

There was discussion about necessary bargaining conditions before negotiations could be progressed besides those already identified. A framework for economic growth and social benefits was identified, but the necessary information on how it might be formulated was uncertain. It seemed clear that in order to see the bargain as having long-term benefit, confidence had to be developed in the MNC's commercial intentions; and again, that would depend on knowledge about the breadth and depth of the MNC's commercial intentions. One participant specified the importance of tech transfer as part of the deal, and the development of feedback loops so that information deficit ongoing would not simply exacerbate misunderstanding and mistrust.

Looking at the consortium scenario, the problem of power imbalances through compounded interests was a recurrent theme. A sense emerged that the bargaining interests of different players were more than they seemed, and how to reveal these was a central information question. Participants wanted to know more about what they were not being told. The experience of other states in dealing with consortium members was suggested as a data source, but problems with confidentiality agreements would arise. In particular, information about the bank's standing within the international financial sector, along with more detail about the terms of the loan and penalties for default were required. The issue of imported labor for the construction company was not considered acceptable because it was not explained beyond skills, and if local labor was not involved there would be no instructional benefit through the exercise. The immigration law implications would lead to a need for "whole of government" information sharing.

The precluded options for power development provoked a need for much more data about the proposed nuclear option, as well as its risks and benefits. Additionally, the rejection of solar options would not be acceptable without some comparative market/environmental analysis.

Particularly, when it came to the push for 5G technology, participants felt totally disadvantaged by knowledge deficit concerning the technology and the implications of coincidental obsolescence. As there was no indication by the consortium of the sustainability of this new technology following its introduction, data about which only the consortium could furnish, participants expressed no position for evaluating cost/benefit. Local business concerns needed development so that they could be put to the consortium proposer for its response, which then would require external evaluation.

When invited to dissect the consortium offer, the fear was expressed that to cherry-pick might mean the loss of desperately needed foreign direct investment. Without environmental impact evaluation for the medium and long term, it was difficult to assess whether the costs attendant on the FDI would outweigh the boost to foreign capital, particularly minus clear capacity building concessions.

The final hypothetical canvassing cash crop diversification also presented regional relations issues. The question of crop security was not addressed and needed to be. However, as with most of the information deficit pertaining to this proposal, there would be a disempowering reliance data sourced from the other bargaining party. This situation emphasized a perennial concern about data validation.

As the arrangement could degenerate into little more than the participant states providing the "farm" for all the offshore commercial benefits, there needed to be information on plans for sustainability, and benefits for the domestic economy. This scenario presented tensions between macro and micro policy desires (diversified cash cropping vs. domestic security and reputational issues), and information was needed in the form of projections on the wider socio-political consequences going forward. Special mention was made about the importance of labor-force benefits, not just through the proposed (but unspecified) R&D injection, but more generally regarding associated agricultural labor transition and mobility. For instance, what would be the concentration on planting/harvesting technology? Participants felt empowered at least to require a detailed business plan from the proposer. Worries about the development of an underground economy in parallel and the exacerbation of already-existing drug problems required thinking through.

## Focus Group 2 – General Observations

At the outset, there was general comment that reflected the concerns of those in the first focus group without then often moving to advance specific information technology solutions. This reluctance could have been a consequence of insufficient clarity that we were not looking just to throw tech at any information deficit. In addition, it reflected the group's belief that data and associated information technology gains its relevance from the questions first asked about need.

The most significant takeaway for the facilitator was the need for a two-pronged approach to information disempowerment, which marries mundane data collection and access devices/routines with capacity enhancement among those who will apply the information to the decision-task. This does not come, originate, or exist as a generalized application, and instead requires purpose of design, modification, and infrastructure support, which we did not get to specify in every hypothetical area of need identified in the first focus group. The main impression for the rapporteur was regular reference to needing to know what the data problem was, which required a data resolution: meaning that both the initial need and whatever data collected may satisfy it, should be clearly specified. Obviously, these observations return to a knowledge gap issue and the requirement for capacity building rather than information tech on its own.

An important qualification about the information empowerment thesis is its present over-emphasis in the current project design on state capacity building. Participants mentioned the not-uncommon situation where a state can use information enhancement for purposes which may advance economic interests at social cost. There was also discussion of the need to ensure information empowerment to the private sector, where trade bargaining and resource retention are matters shared between the state and commerce. Finally, in order that the use of data for trading and resource retention decision-making is for social good, any information enhancement project should not leave out civil society if it is to have the capacity of keeping the other two market players accountable.

Following on from identifying need and sourcing data, discussions included validation and evaluation approaches. With diversity in sources of data, how does one deal with bias? Questions were raised about maintaining the currency and value of data. Original difficulties with knowing what questions to ask might translate into not knowing in what format to employ, store, and order data, or even what the data can accomplish. Added to these are problems of granularity, and the potentially high costs of storage and analysis systems. Connected were worries about giving more data back to companies through information loops and thereby entrenching the information asymmetries in bargaining relativity even further.

On building tech capacity domestically, the data market in the hypotheticals is situated now around identifiable information management needs, so perhaps we are moving into a world where start-ups can be generated without too much capacity required, and these innovators could contribute home-grown information enhancement technologies. How hard would that be? Is it possible to seed something like this? The simplest sustainable solution for information enhancement is to raise capacity within these vulnerable economies to create purpose-designed tech solutions.

On the standardization of data collection vs. having a problem to resolve and then standardizing data afterwards, some participants emphasized "big is best" – the more data you have the better the standardization will be, as well as its application to progressive information needs. A basic observation was made about the utility of producing mundane data from documenting various stages of the supply chain/trade decision-making/auditing processes. Being involved in data production internal to the decision process enables participants to feel that they own that data and understand it better.

## Policy Reflections from the Focus Group Deliberations

Information asymmetries on which the project is based:

- Relationship trade bargainer with external partners
  - Necessary to have knowledge about possible trading partners consequential of any trade negotiation – building contacts with such contacts
  - Knowledge about external companies that are offering trade relationships

- Domestic market information gaps
  - Knowledge about demographics of certain markets: e.g., different fishing practices, how fish stocks may be implicated by trade negotiations
  - Knowledge of existing needs of businesses and commercial relationships with or without trade bargain

- Knowledge gaps in technology
  - Emerging technology: target vulnerable economies may be hampered by limited information about these technologies including about servicing and maintenance in the long run. In turn, this traps technology recipients into a relationship with an external organization in the long run which might compromise the sustainability of the suggested tech aid and increase information dependencies

Capacity building considerations regarding asymmetries and dependencies:

- Capacity building to address knowledge gaps in technology and to enable maintenance of technology in the long-run, or to shift away from an over-reliance from single service providers
  - Work to address dependencies concerning data sources, data integrity, and the accountability of tech development. If people do not know what kinds of questions to ask, it will have consequences for data collection, cleaning, processing, and AI-products chosen and employed. In addition,

haphazard or careless data collection may entrench information asymmetries with external data collectors even further and lead to greater inequalities
  - A working knowledge of technology would aid clarification of when not to use technology
- For trade negotiators (and the wider associated organizations) working in targeted vulnerable economies that still have limited digitalization and technological capacity, consider steps that would make the collection of mundane data more efficient (and not technology dependent) in the near or mid-term.

## Before the injection of AI-assisted information technology

- At the initiation of the project, an intensive needs analysis must be commenced, which is grounded in developing skills around what questions to ask about information deficit, that then will translate into learning about what format to store and order data, and what data can accomplish in trading negotiations and domestic market sustainability.
- Capacity building within the target vulnerable economies will help the identification of major decision sites and bargaining points in the entire supply chain so that negotiators will see where information deficit needs to be addressed.
- International organizations can assist in capacity building as they do not have commitments to either side of any trade bargain. However, due to the lack of relationships between the target vulnerable economy and IOs, consequent on the absence of commercial trading markets on which they may have advised, as well as failure by the target economy in the past to implement international standards, these relationships may need to be project-specific.
- Associated with assistance from international organizations, the target vulnerable economy needs to have access to knowledge in the public domain about natural resources and the demographics of different harvesting practices, and how the relative sustainability of natural resource stocks is impacted

by trading and domestic market decisions. This information access could be provided through aid agencies connection with national scientific repositories and regional data bases.

- Target economies must be trained in all areas of government information retention, usage, and exchange, rather than operate with information locked in certain ministries.
- While information technologies are a priority for advanced consumerist economies, this is not the experience in target vulnerable economies. Therefore, prior to the roll-out of such technologies, sponsors and providers should supplement local limited information about servicing a technology, and the dangers inherent in locking into a provider/client relationship in the long-run.
- Along with technological "needs and potential" training, target vulnerable economies and countries in their region will have limited basic market information to do cost-benefit analysis. However, these economies can supplement this information if provided by aid agencies and IOs with essential local knowledge of social/political contexts in which information is best contextualized.
- Through any phase of externally supported capacity building, there is a need to ensure civil society remains in the loop – to understand business needs on the ground.

## At the introduction of AI-assisted information technology, and following

- Product sustainability is essential and takes certain crucial forms that must be ensured: data sources – who is collecting data and originally for what purposes; data integrity and validation – how is information to be accredited and verified ongoing?; accountability – ensuring that civil society is informed about the type of information that is being collected and provided to governments (particularly important when local farms and fisheries are part of the data production chain); technical sustainability of a technical product – who maintains it? These issues require allied services from sponsors, providers, advisers, and locally trained experts.
- Mission creep: if we want to avoid the monetization of technical applications, developers need a clear and disciplined purpose which is struck in agreement with the local end users.
- At the time of introduction, there should be stimulated public debate about intentions around information access and use. Civil society will then be involved in a holistic and integrated approach to data empowerment.
- As a condition of the technology contract, a home-grown sector development and training in technology development in country must be offered. This could be coordinated and stimulated by a developer-centric branch of government.
- Recognize the importance and resourcing of internet penetration into social networks within the economy, particularly those that provide rich sources of resource and market data.
- Recognize the necessity for introduced technology to be affordable, maintainable, and anti-obsolescent.
- Efforts at standardization in the collection of data from multiple sources, which may then enable more actions on the data to be taken in the analysis phase. This endeavor will include leveraging existing collection methods active and accessible prior to the information roll-out. Identification of mundane data built into the consciousness of people who are currently promoting trade and measuring markets (internal and external to the economy).
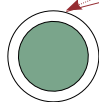
## Concluding Reflections

The rhetoric of what AI can and cannot do continues to be shrouded in mysticism, technological elitism, post-colonial reticence, and just the type of knowledge/power differentials that this project set out to address (de Saint Laurent, 2018). Along with confusion in language and application, AI-assisted decision-making technologies largely remain the province of powerful economies and as such increase their advantages in trade bargaining. In this paper, we have laid out a set of selection-criteria for identifying AI-enabled technologies applicable to assist and restore some balance to trade negotiations by opening up pathways of information and analysis. The criteria for selecting technologies lead on to broader proposals for information access and analysis that will need to be thoroughly contextualized for each vulnerable economy eventually selected for the real-world project. While the method we have piloted has proved useful for enabling the articulation of known unknown factors influencing the relationship between information access and trading power, and these in turn will better enable sustainable trade negotiations (through power dispersal and sharper market discrimination with more information); another key contribution this multi-disciplinary method offers is the identification of pre-existing unknown knowns (such as alternative sites for data access and/or management, and contextual variables which impact information availability and access, irrespective of AI-assisted technologies).

Acknowledging that AI-assisted information technology access alone will not level the trade bargaining horizon or open up understandings of domestic market sustainability, the scoping and solution exercises suggested some essential pre-conditions: (1) participants in the first group were advantaged as they were familiar with, or had a working knowledge of, current ML technologies; and (2) technical experts in the second had a similar working knowledge of trade bargaining theory, so as to prevent technological solutionism that ignored important social, political, and economic contextual

variables influencing capacities to seek out and understand information asymmetries. Some working technical knowledge connected with contextual sensitivities would ensure that people from both groups are speaking in, not the same language, but unfamiliar ground; and that the articulation of need and sustainability of technological solutions can be as precise as possible. In addition to domestic education and training programs, international organizations have a larger role to play in addressing and reversing the knowledge deficits of technologies in trading and domestic market situations as yet deprived of AI-assisted information pathways. International organizations such as UNCTAD and/or the WTO should thus formulate education policies crafted to enable such productive forms of knowledge exchanges to be initiated before the commensuration of the first scoping exercise. More research can be done here to determine productive forms of such exchanges, and their trajectories. Private sector participants looking for a more resilient global trading and sustainable market future also have a role to play here, as do the large information platform providers in helping to achieve the ESCAP sustainability goals.

The potential for enhancing regional cooperation – in addition to the identification of data pathways – through this method can also serve as a route towards increasing industry standardization and state-to-state data flows, particularly where regional sustainability issues are in the trading conversation. Empowerment approaches beyond nation-state priorities are more likely to achieve scalable deployment and interoperability across countries and can be significantly aided by international coordination bodies such as UNCTAD and/or the WTO working together with standard setting bodies, such as the IEEE[32]. At this policy level, trading benefit will be viewed as more than only a national concern. Regional approaches to information empowerment and technological capacity building are a realistic recognition that the information which may assist vulnerable economies often knows no jurisdictional boundaries.

---

32. For example, the IEEE's Data Trading System Initiative. https://standards.ieee.org/industry-connections/datatradingsystem.html

## References

de Saint Laurent, C. (2018). In Defence of Machine Learning: Debunking the Myths of Artificial Intelligence. Europe's Journal of Psychology, 14(4), 734-737.

Dikowitz, S. (2014). World Peace Through World Trade. Retrieved from Hinrich Foundation: https://hinrichfoundation.com/blog/global-trade-world-peace-through-world-trade/

Findlay, M. (2017). Law's Regulatory Relevance? Property, Power and Market Economies.

Fineman, M. A. (2019). Vulnerability and Social Justice. Valparaiso University Law Review 53.

Grcic, J. (2007). Hobbes and Rawls on Political Power. Ethics & Politics. Retrieved from https://core.ac.uk/download/pdf/41174053.pdf

Grossman, W. L. (1935). The Legal Philosophy of Roscoe Pound. Yale Law Review, 605-618.

Hardt, M., & Negri, A. (2001). Empire. Cambridge: Harvard University Press.

Humbach, J. A. (2017). Property as Prophesy: Legal Realism and the Indeterminacy of Ownership. Case Western Reserve Journal of International Law, 211-225.

IT Pro team. (2020, March 2). How to measure data quality. Retrieved from ITPro.: https://www.itpro.co.uk/business-intelligence-bi/29773/how-to-measure-data-quality

Reichel, A. (2018). De-growth and Free Trade. Retrieved from https://www.andrereichel.de/2016/10/18/degrowth-and-free-trade/

Stiglitz, J. (2002a). Globalization and its Discontents. New York: Penguin, 107.

Stiglitz, J. (2002b). Globalization and its Discontents. New York: Penguin, 246.

Susaria, A. (2019). The New Digital Divide is People who Opt out of Algorithms and People who. Retrieved from The Telegraph: https://www.thetelegraph.com/news/article/The-new-digital-divide-is-between-people-who-opt-13773963.php

UNCTAD. (2015). General Assembly: Resolution adopted by the General Assembly on 19 December 2014. United Nations. Retrieved from https://unctad.org/en/PublicationsLibrary/ares69d204_en.pdf

UNTCAD. (2018, October 15). Small economies welcome AI-enabled trade tool but worries remain. Retrieved from UNTCAD: https://unctad.org/en/pages/newsdetails.aspx?OriginalVersionID=1881

UNCTAD. (2019). Digital Economy Report 2019: Value Creation and Capture: Implications for Developing Countries. United Nation. Retrieved from https://unctad.org/en/PublicationsLibrary/der2019_en.pdf

UNCTAD. (2019, July 19). Fairer trade can strike a blow against rising inequality. Retrieved from UNCTAD: https://unctad.org/en/pages/newsdetails.aspx?OriginalVersionID=2154

Xiang, A., & Raji, I. D. (2019). On the Legal Compatibility of Fairness Definitions. Retrieved from https://arxiv.org/abs/1912.00761

# Appendix 1: Hypotheticals

## Instructions

Remember your character and your professional location. Reflect on the facts of the following hypotheticals from the perspective of your character and what you understand to be the "knowledge capacity" of trade and sustainability decision-making in your professional location.

Read the following hypotheticals and imagine you are required to participate and to make decisions as instructed with the information provided. At each nominated decision stage, think about what additional information might be useful in making a more effective choice as the factors of the bargain/retention policy are set out.

Clearly, it is difficult to speculate on what you do not know or what is being withheld from you. In this context, common sense as well as experience are useful measures in determining how your decision/bargain would be more empowered through the information available to you. One way of approaching this is to think about the issue/problem that you are confronting, where might be a source of information you currently do not possess, and the form that information might take.

Finally, you are not entirely unfamiliar with information technology. Even though official data, retrieval, and analysis capacity in your professional context is limited, you have a sense of what technological enhancements and information databases those in better resourced administrations and commercial arrangements can access and use to their benefit (and perhaps your detriment). Therefore, you are concerned with information deficit and what information access might enable. You are also interested in how information can be analyzed and applied to make your professional experience more efficient and sustainable.

## Hypothetical 1

A large multi-national corporation has commenced discussion with your government to have access to fishing grounds in your territorial waters. Due to the tariff war between several other much larger fishing nations, the price of fish products has grown incrementally in the last economic quarter. The multi-national is also attracted to a trading arrangement with you because your national regulation of fishing practice is neither detailed nor unduly restrictive. In fact, global fishing quotas have largely had little impact on your domestic fishing practice because of its up-until-now subsistence format.

The multi-national has not divulged its intended market for the fish products it would acquire from your waters, but you have some general intelligence that Japan would be a principal third party trader. In Japan, you are aware that the consumer appetite for one particular fish product which is abundant in your waters is high, and prices that can be fetched seem to you to be extraordinary. You have no developed trade arrangement with Japan and you have no detailed understanding of their fish product consumer markets.

The multi-national has also expressed interest in using local labor, the price of which is under-valued due to limited local employment opportunities in the sector. In preliminary meetings, the multi-national has talked of building canning factories for fish processing in two of your major ports where female unemployment is particularly high.

Fish are a dietary staple for many of your citizens living in coastal regions, who practice small scale, indigenous fishing practices. Your fisheries and wildlife department has not done any study on the fish stocks in your territorial waters or on the impact of large-scale commercial fishing on these stocks. You do not have up-to-date information on the multi-national's practices in the harvesting and use of natural resources. In these negotiations, you would be dealing with a subsidiary of the larger multi-national set up specifically for this trading exercise and registered in the Republic of Ireland for beneficial taxation concessions.

You have been asked:

a) To further the preliminary negotiations with the multi-national;
b) To oversee an environmental impact assessment of the proposal;
c) To draft conditions under which specific trade negotiations might be structured;
d) To address concerns from local indigenous fishing communities.

## Hypothetical 2

A consortium of foreign investors has approached your government with the intention of structuring and implementing some foreign direct investment (FDI) infra-structure projects in your country. The consortium consists of a major Chinese banking group, an international construction company, a major power generator, and a telecommunications provider. The types of projects being discussed are very attractive to your under-capitalized transport and communications sector.

A condition of the foreign direct investment portfolio is that your government signs up to various loan agreements offered by the Chinese bank. As a condition of the loans, your government will agree to having any disputes arising between your state and the consortium arbitrated in China under Chinese commercial law.

The international construction company will design and build a new dam over a large natural river system. Water resources are a major concern for your country. Because of what they refer to as 'technology considerations', the construction company intends only to use its own imported labor.

Your state is in desperate need of power generating facilities. The major power generator in the consortium is happy to finance the construction and operation of a nuclear power plant within your territory, provided that you allow half of the power generated in that grid to be independently traded by the consortium into neighboring states. In addition, the consortium wants your government to cease discussions with another neighbor states for the shared construction of wind farms on your border.

The telecommunications provider will invest in 5G technologies throughout your state. Most of your communication capacity at present is not fully 4G compliant. There have been concerns expressed in your business community that such a rapid convergence into 5G might produce significant secondary costs through unnecessary technological obsolescence. Furthermore, talk from the telecommunications about linking your 5G capacity to developments in the Internet of Things (IoT) in China, seem obscure and unclear.

You have been asked:

a) To further the preliminary negotiations with the consortium;
b) To oversee an economy-wide evaluation of the impact of the proposed FDI;
c) To draft conditions under which specific investment negotiations might be structured;
d) To address concerns from local businesses such as the domestic power provider, domestic telcos, and local trade unions regarding medium-term sustainability issues.

## Hypothetical 3

In an effort to improve your trade imbalance, your government over recent decades has implemented an agricultural policy of transition from subsistence to cash cropping. In particular, palm oil plantations have been incentivized and major regional companies have invested in concessions for palm oil production. A political consequence has been push-back from smaller farmers who are unable to match the economies of scale of the bigger plantations. To confront this resistance, the government has operated a subsidy system to encourage small farmers to cash crop, and to compensate for their market disadvantage.

Both the bigger producers and the small farmers employ slash-and-burn clearing techniques, which has caused air pollution with associated damage to the health of the domestic population and neighboring states.

The government is worried about its growing dependence on a single export crop, when global market vulnerability is difficult to predict. Entrepreneurs from Canada, which recently legalized the growing and use of marijuana, are in discussions with your government to invest in major hemp farms in your country for export back to Canada and California, where they say the market is expanding. Governments in your region with tough anti-drug laws have lobbied your government against the initiative. Marijuana is currently a prescribed drug in your jurisdiction, but popular opinion would be tolerant of decriminalization for medical and economic reasons.

The Canadian investors have also indicated – to improve the attractiveness of their agricultural intentions – to bring with them a significant research and development investment that could stimulate the growth of a generic drug industry in your country; namely, processing the medical constituents of marijuana. This industry would, they say, offers employment mobility for semi-skilled workers currently occupied in low-paid sweat shop garment-making, which is another diminishing domestic export industry here.

You have been asked:

a) To further the preliminary negotiations with the Canadian investors;
b) To oversee a comparative environmental impact assessment of the proposal relative to existing cash cropping practices;
c) To draft conditions under which investment negotiations might be structured;
d) To address concerns on the relationship between trade and regional foreign policy.

# Governing Data-driven Innovation for Sustainability: Opportunities and Challenges of Regulatory Sandboxes for Smart Cities
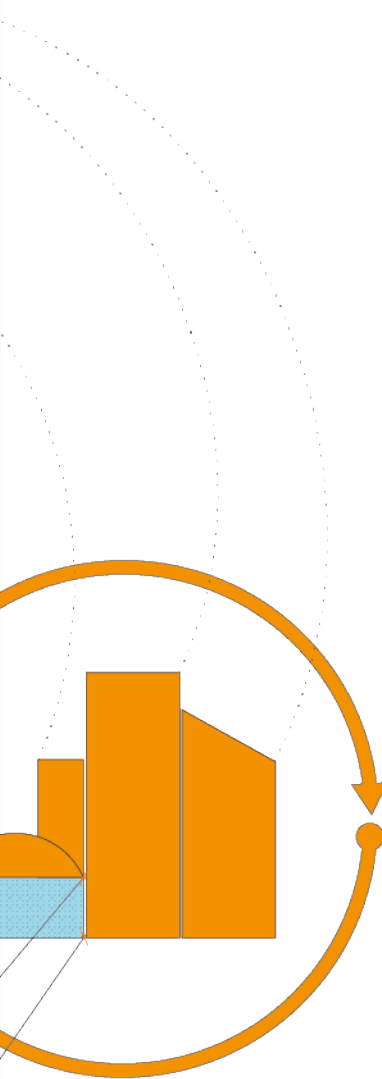
**Masaru Yarime**[1]
Division of Public Policy,
The Hong Kong University of
Science and Technology
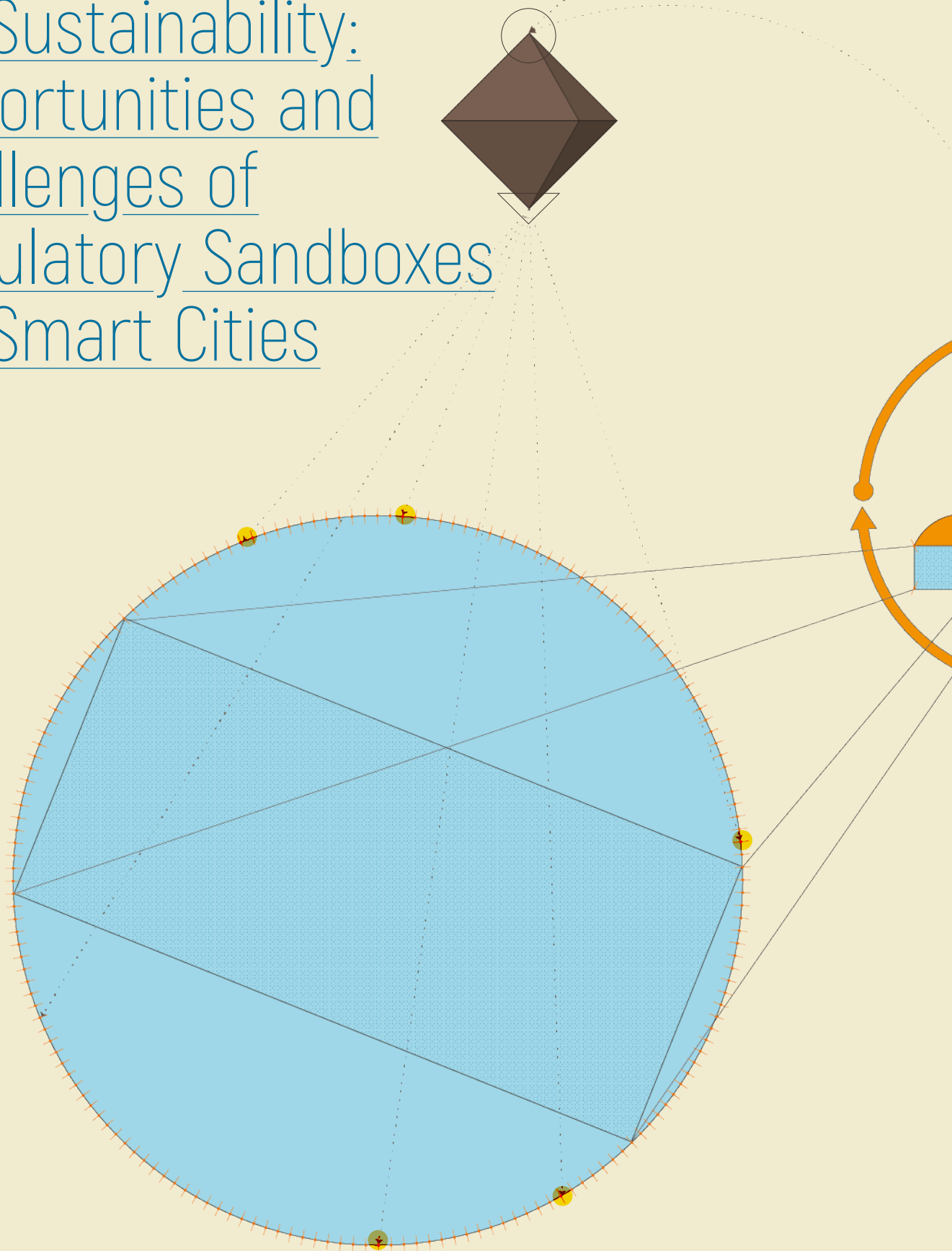
## Abstract

Data-driven innovation plays a crucial role in tackling sustainability issues. Governing data-driven innovation is a critical challenge in the context of accelerating technological progress and deepening interconnection and interdependence. AI-based innovation becomes robust by involving the stakeholders who will interact with the technology early in development, obtaining a deep understanding of their needs, expectations, values, and preferences, and testing ideas and prototypes with them throughout the entire process. The approach of regulatory sandboxes will particularly play an essential role in governing data-driven innovation in smart cities, which inevitably faces a difficult challenge of collecting, sharing, and using various kinds of data for innovation while addressing societal concerns about privacy and security. How regulatory sandboxes are designed and implemented can be locally adjusted, based on the specificities of the economic and social conditions and contexts, to maximize the effect of learning through trial and error. Regulatory sandboxes need to be both flexible to accommodate the uncertainties of innovation, and precise enough to impose society's preferences on emerging innovation, functioning as a nexus of top-down strategic planning and bottom-up entrepreneurial initiatives. Data governance is critical to maximizing the potential of data-driven innovation while minimizing risks to individuals and communities. With data trusts, the organizations that collect and hold data permit an independent institution to make decisions about who has access to data under what conditions, how that data is used and shared and for what purposes, and who can benefit from it. Alternatively, a data linkage platform can facilitate close coordination between the various services provided and the data stored in a distributed manner, without maintaining an extensive central database. The data governance systems of smart cities should be open, transparent, and inclusive. As the provision of personal data would require the consent of people, it needs to be clear and transparent to relevant stakeholders how decisions can be made in procedures concerning the use of personal data for public purposes. The process of building a consensus among residents needs to be well-integrated into the planning of smart cities, with the methodologies and procedures for consensus-building specified and institutionalized in an open and inclusive manner. It is also essential to respect the rights of those residents who do not want to participate in the data governance scheme of smart cities. As APIs play a crucial role in facilitating interoperability and data flow in smart cities, open APIs will facilitate the efficient connection of various kinds of data and sophisticated services. International cooperation will be critically important to develop common policy frameworks and guidelines for facilitating open data flow while maintaining public trust among smart cities across the globe.

## Introduction

Data-driven innovation plays a crucial role in tackling sustainability challenges such as reducing air pollution, increasing energy efficiency, eliminating traffic congestion, improving public health, and maintaining resilience to accidents and natural disasters (Yarime, 2017). These multifaceted challenges, which are interconnected and interdependent in complex ways, require the effective use of various kinds of data concerning environmental, economic, social, and technological aspects that are increasingly available through sophisticated equipment and devices in smart cities. Innovation based on artificial intelligence (AI) can make the best use of these data to accelerate learning and improve performance. It is of critical importance to establish adaptive governance systems that allow experimentation and flexibility to deal with the uncertainty and unpredictability of technological change, while addressing societal concerns such as security and privacy incorporating local contexts and conditions. Novel forms of technology governance, such as testbeds, living laboratories, and regulatory sandboxes, are required for policymakers to address the evolving nature of data-based innovation.

In this paper, we examine key opportunities and challenges in the governance of data-driven innovation in the context of smart cities. First, we discuss the major characteristics of data-driven innovation and highlight the importance of learning and adaptation through the actual use of technologies in real situations. Next, we examine the approach of regulatory sandboxes to facilitate innovation by taking previous examples of introducing them to the field of finance and other sectors with their experiences and implications. Then we consider emerging cases of applying regulatory sandboxes to stimulate novel technologies utilizing AI in cyber-physical systems such as drones, autonomous vehicles, and smart cities. Finally, we discuss critical challenges in designing and implementing regulatory sandboxes for AI-based innovation, with a particular focus on

data governance. Implications are explored for data governance to promote the collection, sharing, and use of data for innovation while taking appropriate measures to address societal concerns, including safety, security, and privacy. Recommendations for policymakers are considered to facilitate the engagement of relevant stakeholders in society so that various kinds of data collected in smart cities are appropriately used to govern innovation based on AI.

## Characteristics of Data-driven Innovation

The emergence of data-driven innovation based on the rapid advancement in the Internet of Things (IoT) and AI creates exciting opportunities as well as considerable challenges in promoting societal benefits while regulating the associated risks. As a vast amount of diverse kinds of data is increasingly available from various sources that were not previously accessible, a wide range of sectors are currently undergoing significant transformation. In energy, smart grid systems lower costs, integrate renewable energies, and balance loads. In transportation, dynamic congestion-charging systems adjust traffic flows and offer incentives to use park-and-ride schemes, depending upon real-time traffic levels and air quality. Car-to-car communication can manage traffic to minimize transit times and emissions, and eliminate road deaths from collisions (Curley, 2016). The speed of technological advancement is accelerating, and those technologies that used to be separate are increasingly interconnected and interdependent with one another, creating a significant degree of uncertainty in their impacts and consequences.

The process of data-driven innovation has three key components: data collection, data analysis, and decision making (Organisation for Economic Co-operation and Development, 2015a). Data-driven innovation critically depends on the efficient and effective collection, exchange, and sharing of large

amounts of high-quality data. New technologies such as drones, IoT, and satellite images can now provide vast amounts of data that were not previously available or accessible before. The big data collected through various sources and challenges are analyzed by applying data science. Sophisticated methodologies and tools are increasingly possible due to the recent technological advancement in AI, particularly the rapid progress in machine learning. For decision making, it is critical to integrate the findings of data analytics with the domain expertise that would be specific to the sector in which you are involved, such as energy, health, or transportation. Increasingly, cyber systems are merging with physical machines and instruments as in manufacturing, and such cyber-physical systems are particularly important in dealing with sustainability issues in the context of smart cities.

Data-driven innovation is accelerated by deriving new and significant insights from the vast amount of data generated during the delivery of services every day. Hence training, the ability to learn from real-world use and experience, and adaptation, the capability to improve the performance, would be key to creating data-driven innovation (Food and Drug Administration, 2019). The development of cyber-physical systems such as smart cities is facilitated through the ready availability of and accessibility to data, as well as its mutual exchange and sharing with stakeholders in different sectors. Unlike the traditional model of innovation, which tends to rely on closed, well-established relationships between enterprises in a specific industry, the new mode of data-driven innovation requires open, dynamic interactions with stakeholders possessing and generating various kinds of data. Close cooperation and collaboration on data become crucial in the innovation process, from the development of novel technologies to deployment through field experimentation and legitimation in society.

There are difficult challenges to policymakers in facilitating data-driven innovation in cyber-physical systems. The speed of technological change of AI is remarkably fast, which has been particularly demonstrated in the case of image recognition (Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg & Fei-Fei, 2015). That leads to remarkable progress in the performance of AI and, at the same time, accompanies a significant degree of uncertainty in consequences and side effects. Various kinds of technologies are increasingly interconnected and interdependent through data exchange and sharing among multiple sectors, such as energy, buildings, transportation, and health. These characteristics make it difficult to explain or understand the process of innovation and contribute to giving rise to a widening gap between technological and institutional changes. It is critical to establish a proper system to govern data-driven innovation in the context of accelerating technological progress and deepening interconnection and interdependence. New policy approaches are required to stimulate data-driven innovation in cyber-physical systems by facilitating coordination and integration of emerging technologies while addressing societal concerns such as safety, security, and privacy.

As the introduction of AI systems is relatively new, our understanding of the behavior of such systems in real-life situations is still minimal. As machines powered by AI increasingly mediate our economic and social interactions, understanding the behavior of AI systems is essential to our ability to control their actions, reap their benefits, and minimize their harms (Rahwan, Cebrian, Obradovich, Bongard, Bonnefon, Breazeal, Crandall, Christakis, Couzin, Jackson, Jennings, Kamar, Kloumann, Larochelle, Lazer, McElreath, Mislove, Parkes, Pentland, Roberts, Shariff, Tenenbaum & Wellman, 2019). AI systems cannot be entirely separate from the underlying data on which they are trained or developed. Hence it is critical to understand how machine behaviors vary with altered environmental inputs, just as biological agents' behaviors vary depending on the environments in which they exist. Our understanding of the behavior of AI-based systems can benefit from an experimental examination of human-machine interactions in real-world settings.

The experience of using an AI system in clinics in Thailand for the detection of diabetic eye disease is one of the few cases that provide valuable lessons and implications (Beede, Baylor, Hersch, Iurchenko, Wilcox, Ruamviboonsuk & Vardoulakis, 2020). While deep learning algorithms promise to improve clinician workflows and patient outcomes, these gains have not been sufficiently demonstrated in real-world clinical settings. The Ministry of Health in Thailand has set a goal to screen 60% of its diabetic population for diabetic retinopathy (DR), which is caused by chronically high blood sugar that damages blood vessels in the retina. Reaching this goal, however, is a challenge due to a shortage of clinical specialists. That limits the ability to screen patients and also creates a treatment backlog for those found to have DR. Thus, nurses conduct DR screenings when patients come in for diabetes check-ups by taking photos of the retina and sending them to an ophthalmologist for review. A deep learning algorithm has been developed to provide an assessment of diabetic retinopathy, avoiding the need to wait weeks for an ophthalmologist to review the retinal images. This algorithm has been shown to have specialist-level accuracy for the detection of referable cases of diabetic retinopathy. Currently, there are no requirements for AI systems to be evaluated through observational clinical studies, nor is it common practice. That is problematic because the success of a deep learning model does not rest solely on its accuracy, but also on its ability to improve patient care.

This experience provides critical recommendations for continued product development and guidance on deploying AI in real-world scenarios (Beede, 2020). The functioning of AI systems in healthcare is affected by workflows, system transparency, and trust, as well as environmental factors such as lighting which vary among clinics and can impact the quality of images. AI systems need to be trained to handle these situations. An AI system might conservatively determine some images having blurs or dark areas to be ungradable because they might obscure critical anatomical features required to provide a definitive result. On the other hand, the gradability of an image may vary depending on a clinician's experience or physical set-up. Any disagreements between the AI system and the clinician can create problems. The research protocol has been subsequently revised, and now eye specialists review such ungradable images alongside the patient's medical records, instead of automatically referring patients with ungradable images to an ophthalmologist. This helped to ensure a referral was necessary and reduced unnecessary travel, missed work, and anxiety about receiving a possible false-positive result. In addition to evaluating the performance, reliability, and clinical safety of an AI system, we also need to consider the human impacts of integrating an AI system into patient care. The AI system could empower nurses to confidently and immediately identify a positive screening, resulting in quicker referrals to an ophthalmologist.

This case highlights that, in addition to the accuracy of the algorithm itself, the interactions between end-users and their environment determine how a new system based on AI will be implemented, which cannot always be controlled through careful planning. Even

when a deep learning system performs a relatively straightforward task, for example, just analyzing retinal images, organizational or socio-environmental factors are likely to impact the performance of the system. Many environmental factors that negatively impact model performance in the real world might be reduced or eliminated by technical measures, such as through lighting adjustments and camera repairs. However, these types of modifications could be costly and even infeasible in low-resource settings, making it even more critical to engage with contextual phenomena from the start. AI-based innovation becomes robust by involving the stakeholders who will interact with the technology early in development, obtaining a deep understanding of their needs, expectations, values, and preferences, and testing ideas and prototypes with them throughout the entire process.

The findings of the actual case of implementing AI-based innovation provide useful implications for technology policy and governance. As policy makers are required to respond to technological change in real-life situations, technology governance becomes an integral part of the innovation process itself to steer emerging technologies towards better collective outcomes. Governments need to anticipate significant changes induced by autonomous vehicles, drone technologies, and widespread IoT solutions, as well as to consider their implications for public policy. AI technologies offer opportunities to improve economic efficiency and quality of life, but they also bring many uncertainties, unintended consequences, and risks. As such, this calls for more anticipatory and participatory modes of governance (OECD, 2018).

Anticipatory governance acts on a variety of inputs to manage emerging knowledge-based technologies and the missions built upon them, while such management is still possible (Guston, 2014). It requires government foresight, engagement, and reflexivity to facilitate public acceptance of new technologies, while at the same time assessing, discussing, and preparing for their intended and unintended economic and societal effects. Anticipatory approaches can help explore, consult widely on, and steer the consequences of innovation at an early stage and incorporate public values and concerns, mitigating potential backlash against technology. Traditional policy tools would not be able to deal with situations where the future direction of technological innovation cannot be determined. In contrast, new policy tools such as regulatory sandboxes emphasize the benefits of environments that facilitate learning to help understand the regulatory implications and responses to emerging technologies. Participatory approaches can provide a wide range of stakeholders, including citizens, with adequate opportunities to appraise and shape technology pathways (OECD, 2018). These practices can help ensure that the goals, values, and concerns of society are continuously enforced in emerging technologies, and shape technological designs and trajectories without unduly constraining innovators. This will contribute to supporting efforts to promote responsible innovation, which has integrated dimensions of anticipation, reflexivity, inclusion, and responsiveness (Stilgoe, Owen & Macnaghten, 2013).

# The Approach of Regulatory Sandboxes

The approach of regulatory sandboxes has recently been proposed to stimulate innovation by allowing experimental trials of novel technologies and systems that cannot currently operate under the existing regulations by specifically designating geographical areas or sectoral domains. Regulatory sandboxes provide a limited form of regulatory waiver or flexibility for firms to test new products or business models with reduced regulatory requirements, while preserving some safeguards to ensure appropriate consumer protection (Organisation for Economic Co-operation and Development, 2019). Potential benefits include facilitating greater data availability, accessibility, and usability for innovators, and reducing the time and cost of getting innovative ideas to market by reducing regulatory constraints and ambiguities (Financial Conduct Authority, 2015). The approach aims to provide a symbiotic environment for innovators to test new technologies and for regulators to understand their implications for industrial innovation and consumer protection. The aim is to help identify and better respond to regulatory breaches by enhancing flexibility and adjustment in regulations, which would be particularly relevant in highly regulated industries, such as the finance, energy, transport, and health sectors.

Regulatory sandboxes have initially been introduced to the financial sector in efforts to encourage fintech by providing a regulatory safe space for innovative financial institutions and activities underpinned by technology (Zetzsche, Buckley, Barberis & Arner, 2017). While the sandbox creates an environment for businesses to test products with less risk of being punished by the regulator for non-compliance, regulators require applicants to incorporate appropriate safeguards to insulate the market from risks of their innovative business. In early 2016, the Financial Conduct Authority (FCA) of the UK initiated a fintech regulatory sandbox to encourage innovation in the field of financial technology. The sandbox aimed to

provide the conditions for businesses to test innovative products and services in a controlled environment without incurring the regulatory consequences of pilot projects (Financial Conduct Authority, 2015). A fintech supervisory sandbox was also launched by the Hong Kong Monetary Authority in September 2016, followed by other fintech sandboxes in Australia, Canada, and Singapore. The concept has also been embraced by a growing number of developing world regulators as well.

There are some lessons learned from the experience of regulatory sandboxes in fintech (Financial Conduct Authority, 2017). Working closely with the FCA has allowed firms to develop their business models with consumers in mind and mitigate risks by implementing appropriate safeguards to prevent harm. A set of standard safeguards have been put in place for all sandbox tests. All firms in the sandbox are required to develop an exit plan to ensure that the test can be terminated whenever it is necessary to stop the potential harm to participating consumers. The sandbox has allowed the agency to work with innovators to build appropriate consumer protection safeguards into new products and services.

The approach of regulatory sandboxes has gone beyond the field of finance and has been applied in other sectors involving cyber-physical systems, which more directly concern safety, human health, and public security. In the energy sector, the Office of Gas and Energy Markets (Ofgem) of the UK started their Innovation Link service in February 2017 as a one-stop shop offering rapid advice on energy regulation to businesses looking to launch new products or business models (Office of Gas and Electricity Markets, 2018a). When regulatory barriers prevent launching a product or service that would benefit consumers, a regulatory sandbox can be granted to enable a trial.

The Energy Market Authority (EMA) in Singapore also launched a regulatory sandbox in October 2017 to encourage experimentation of new products and services in the electricity and gas sectors (Energy Market Authority, 2017). EMA, as the industry regulator, assesses the impact of new products and services before deciding on the appropriate regulatory treatment. Innovators submit their ideas to EMA for testing, and a successful application allows the plan to be applied in the market while being subject to relaxed regulatory requirements. Safeguards such as limiting the duration of the trial or the maximum number of consumers can be introduced to minimize risks to consumers and industry. The evaluation criteria when applying for the regulatory sandbox include using technologies or products in an innovative way, addressing a problem or bringing benefits to consumers or the energy sector, requiring some changes to existing rules, and having assessed and mitigated foreseeable risks. The regulatory sandbox complements ongoing energy research and development (R&D) initiatives by providing a platform for R&D projects to be tested on a broader scale in the country.

The experience of introducing regulatory sandboxes to the energy sector offers a number of lessons and implications. Ofgem's officials spent time talking to innovators to understand their business and to locate and interpret the rules that affected them. Through an iterative process, they effectively worked with innovators to co-create feasible sandbox trials (Office of Gas and Electricity Markets, 2018b). It was not always clear to innovators what they could or could not do, nor always easy for them to find rules or interpret them. Hence advice from the agency helped the innovators figure out which regulations would be relevant for their technologies or services. Sometimes proposals were not allowed for trials, as some institutional requirements, including industry norms, systems, charging arrangements, codes, and licenses, became obstacles. While the sandbox was introduced to facilitate time-limited trials with the

temporary relaxation of rules, most innovators would like to continue to operate after the test and to see the experience of regulatory sandboxes used to change the existing policies and regulations.

The approach of regulatory sandboxes can play an essential role in governing data-driven innovation, which inevitably faces a difficult challenge of collecting, sharing, and using various kinds of data for innovation while addressing societal concerns about privacy and security. The Information Commissioner's Office (ICO) in the UK has recently introduced a regulatory sandbox that is designed to support start-ups, SMEs, and large organizations across private, public, and voluntary sectors. The condition is that they use personal data to develop products and services which are innovative and have demonstrable public benefits (Information Commissioner's Office, 2019). The regulatory sandbox enables participants to consider how they use personal data in their projects, as well as provides some comfort from enforcement action and increases public reassurance that innovative products and services are not in breach of data protection legislation. As these products and services are considered to be on the cutting edge of innovation and operating in particularly challenging areas of data protection, there is a significant extent of uncertainty about adequately complying with the relevant regulations. Participants in the regulatory sandbox can become use cases, and, subsequently, the ICO would be able to revise public guidance and provide necessary resources for compliance.

An important issue in designing and implementing regulatory sandboxes is how to manage regulatory arbitrage. Regulatory sandboxes aim to stimulate innovation by relaxing relevant regulations so that entrepreneurs can experiment with novel technologies without being constrained too much by the existing regulatory environment. This creates opportunities for regulatory arbitrage, which refers collectively to the strategies that can be used to achieve an economically equivalent outcome to a regulated

activity while avoiding the legal constraints (Fleischer, 2010). It is a legal planning technique used to avoid regulatory requirements such as taxes, accounting rules, securities disclosure, with other requirements such as safety and privacy also possibly being included. Jurisdictionally speaking, regulatory arbitrage means that a firm chooses a location where a more favorable regulatory treatment is available to its business activities (Allen, 2019). While national borders do not constrain the development and deployment of AI-based products and services, regulatory sandboxes have only been created at national or sub-national levels. This discrepancy can lead to what is known as the race to the bottom, a phenomenon where jurisdictions compete to lower their regulatory standards in order to attract innovative companies, which could potentially result in negative consequences on consumer protection with regard to safety and privacy.

The challenges of regulatory arbitrage and the race to the bottom can be tackled if the regulators in different locations can coordinate with one other to share the information necessary to formulate appropriate policy measures and commit to agreements to apply consistently high regulatory standards (Allen, 2020). Regulators, however, have their specific policy preferences and strong incentives to keep information within individual regulatory sandboxes, rather than share it with other sandboxes in different locations. Social license and the bundling of laws and resources could work as constraining forces on regulatory arbitrage (Pollman, 2019). Aggressive regulatory arbitrage can erode social license and create a costly environment for sustainable operation, especially when social costs are widely recognized in the community. Also, as an opportunity for regulatory arbitrage would arise not in isolation but within a system of laws, and in light of other considerations such as investment capital and workforce talent, the bundling of relevant laws and regulations would leave less room available for regulatory arbitrage. If the existing laws create a regulatory environment that is prohibitive to a particular type of innovation, companies may try to

focus on changing the legal environment rather than merely arbitrage regulatory differences. A complex set of factors and considerations would influence decisions about regulatory arbitrage, which includes transparency of information to the public and the ability of a company to mobilize its resources for regulatory change.

Moving in a more positive direction, an increasing number of enterprises actually try to advance innovative technologies by strategically taking regulatory arbitrage. One example is Cyberdyne, a Japanese company that developed a medical and healthcare robot, HAL (Ikeda and Iizuka, 2019). Under the Japanese product classification system, HAL could be categorized as a medical device or an assistive device, each of which would be regulated by different institutions. Although the company initially planned to commercialize the robot as a medical device with public medical insurance coverage, that required the product to comply with rigorous medical safety regulations with clinical trials. Considering the regulatory environment, Cyberdyne first chose to commercialize HAL as an assistive device, which usually requires proof of safety, certified by a third party on voluntary terms. The Robot Safety Centre, a public institution located in the Tsukuba International Strategic Zone, Tokku, supported the company to conduct the necessary testing and produce evidence for proof of safety. During this process the company was able to accumulate experiences to improve the product, which was eventually certified by the Japan Quality Assurance Organization and commercialized as an assistive device.

On the other hand, Cyberdyne chose to commercialize HAL as a medical device in Germany first (Iizuka and Ikeda, 2019). From the beginning there was an expectation that it would take a long time to receive an approval from the Ministry of Health, Labour and Welfare (MHLW) in Japan because there was no precedent product similar to the new robot. In Germany, in contrast, a new health device like HAL

is categorized solely as a medical device strictly by its function regardless of its risk levels on safety. As the review of medical devices is certified by a private certification body, the procedure is codified, open, and transparent, and the time required for approval of new medical devices is substantially less than in Japan. HAL has been certified as a medical device in Germany and subsequently commercialized in Europe. After that, the robot was approved by the Pharmaceuticals and Medical Devices Agency (PMDA) in Japan and commercialized with public insurance coverage.

At the same time, Cyberdyne also engaged in developing ISO standards for the safety of personal care robots including healthcare robots (Iizuka & Ikeda, 2019). As there had not been robots like HAL before, there was no regulation in place to protect users, and international standards were considered to be crucial for establishing confidence in these products. Also, while these new standards can guarantee the company an early-mover advantage with the global recognition of its brand, they level the playing field for new entrants to the emerging industry. As Cyberdyne was already developing personal care robotics and was experimenting with prototype safety measures, a set of evidence created during this process became a basis to establish ISO standards on robotics safety.

This case demonstrates a possibility that regulatory arbitrage can actually function to promote innovation. As a start-up with limited resources, Cyberdyne did not attempt to directly influence the relevant regulations. The company instead tried to cope with the regulatory obstacle by commercializing the new robot in the domestic market as an assistive device first and further developing the technology as a medical device overseas. The company also participated in setting up the institutional environment in which the new product is recognized properly. Hence regulatory arbitrage can also mean that enterprises strategically take advantage of differences in regulatory systems to develop and commercialize innovative products while contributing to establishing institutions to facilitate market creation.

## Cases of Regulatory Sandboxes for AI-based Innovation

For smart city development, demonstration projects play an increasingly crucial role in testing novel technologies and raising awareness among the general public. These projects are mainly aimed at examining promising but unproven technologies concerning various aspects of cities, including energy, transportation, buildings, health, environment, and infrastructure. Existing policies and regulations, however, may not necessarily be able to properly deal with certain unexpected novel features of technologies. Hence entrepreneurs and innovators would have difficulties in conducting field testing of emerging technologies on the ground, particularly when other stakeholders, including local communities and residents, are involved. Regulatory sandboxes can relax or adjust some of the relevant regulations so that these new technologies can be tested for actual adoption and use. How regulatory sandboxed are designed and implemented can be locally adjusted, based on the specificities of the economic and social conditions and contexts, to maximize the effect of learning through trial and error. Various types of new promising technologies can be verified, adopted, and integrated, effectively improving technological performance, reliability, and integration, as well as contributing to cost reduction.
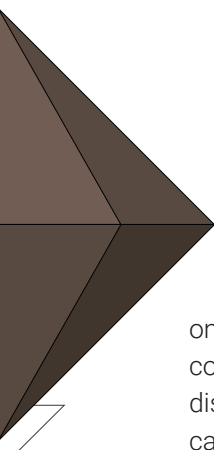
In particular, regulatory sandboxes can improve the understanding of how AI systems may react in specific contexts and satisfy human needs. As AI-based innovation involves rapid technological change, uncertain market development, and diverse social norms, there are many economic, ethical, and legal issues comprised of various interests and preferences. It is necessary to have a regulatory framework that is flexible enough to accommodate the uncertainties of innovation and, at the same time, clear enough to impose society's preferences on emerging innovation. This requires a specific form of governance that incorporates both elements of top-

down legal framing and bottom-up empowerment of individual actors (Pagallo, Aurucci, Casanovas, Chatila, Chazerand, Dignum, Luetge, Madelin, Schafer & Valcke, 2019). Regulatory sandboxes can function as a nexus of top-down strategic planning and bottom-up entrepreneurial initiatives.

The current regulations in the fields of autonomous vehicles, drones, and medical devices show that rules on AI are significantly dependent upon the context of locations and sectors (Pagallo, Aurucci, Casanovas, Chatila, Chazerand, Dignum, Luetge, Madelin, Schafer & Valcke, 2019). In the case of the EU, for example, in addition to the rules on data protection, the testing and use of self-driving cars needs to comply with a complex legal network involving three directives and one regulation: Council Directive 85/374/EEC on the approximation of the laws, regulations, and administrative provisions of the Member States concerning liability for defective products; Directive 1999/44/EC on certain aspects of the sale of consumer goods and associated guarantee, such as repair and replacement, and price reduction and termination; Directive 2009/103/EC relating to insurance against civil liability in respect of the use of motor vehicles, and the enforcement of the obligation to insure against such liability; and Regulation 2018/858 on the approval and market surveillance of motor vehicles and their trailers, and of systems, components, and separate technical units intended for such vehicles. The testing and use of drones requires compliance with one regulation, Regulation (EU) 2018/1139 on common rules in the field of civil aviation and establishing a European Aviation Safety Agency, and two European Commission implementing and delegated acts, Delegated Regulation 2019/945 and the Implementing Regulation 2019/947, in addition to several opinions and guidelines of the European Aviation Safety Agency (EASA). Medical devices based on AI need to deal with contractual and tort liability in national regulations of the EU member states.

Given the rapid progress and unpredictable evolution of AI-based innovation, some countries have established special deregulated zones as living labs to allow testing and experimentation of new technologies in actual fields. In Japan, the National Strategic Special Zones system was introduced in 2013 to enhance economic growth by implementing regulatory reforms. So far, ten areas have been designated as special zones, and more than 60 reforms have been realized, with over 350 projects currently ongoing as a result of these regulatory reforms (Secretariat for the Promotion of Regional Development, 2019). In these special zones, regulatory exceptions have been introduced without amending the laws by taking into account specific local circumstances, and municipalities and private companies have proposed voluntary plans. Specifically targeting self-driving vehicles, in October 2017, the government introduced the National Strategic Special Zones for Level 4 Automated Vehicles Deployment Project on public roads. With the aim of establishing social and legal systems for future technological development, public road safety demonstration experiments were conducted. Based on the experience of building these special zones, the Japanese government initiated a new framework for regulatory sandboxes in March 2018, covering financial services, healthcare industry, mobility, and transportation.

In Singapore, the Road Traffic Act was amended in February 2017 to recognize that a motor vehicle need not have a human driver. The Minister for Transport is able to create new rules on trials of autonomous vehicles, acquire the data from the trials, and set standards for autonomous vehicle designs (Taeihagh & Lim, 2019). A five-year regulatory sandbox was created to ensure that innovation is not stifled, and the government intends to enact further legislation in the future. Autonomous vehicles must pass safety assessments, robust plans for accident mitigation must be developed before road testing, and the default requirement for a human driver can be waived

once the autonomous vehicle demonstrates sufficient competency to the Land Transport Authority. After displaying higher competencies, autonomous vehicles can undergo trials on increasingly complex roads.

In 2017, the United States Federal Aviation Administration (FAA) launched the Unmanned Aircraft System (UAS) Integration Pilot Program (IPP), with fixed-term regulatory exemptions and adaptive regulations, to test the safe application of drones (Federal Aviation Administration, 2019). The program has helped the Department of Transportation and FAA develop new rules that support more complex low-altitude operations by addressing security and privacy risks and accelerating the approval of operations that currently require special authorizations. Ten public-private partnerships have been chosen to test the use of unmanned aerial vehicles (UAV), drones, in potentially useful ways that are currently illegal under federal law without a waiver (Boyd, 2018). The program encouraged applicants to submit proposals for test cases that would obtain data that could be applied to broader use cases, with the understanding that the Department of Transportation and FAA would waive certain restrictions to make these projects viable. The IPP Lead Participants are evaluating a host of operational concepts, including night operations, flights over people and beyond the pilot's line of sight, package delivery, detect-and-avoid technologies, and the reliability and security of data links between pilot and aircraft, with potential opportunities for application in commerce, photography, emergency management, agricultural support, and infrastructure inspections.

In Germany, the energy sector is emphasized to encourage innovative solutions for a future energy system based on renewable energy and higher energy efficiency through digitalization. The Economic Affairs Ministry has set up a large-scale regulatory sandbox entitled Smart Energy Showcases – Digital Agenda for the Energy Transition (SINTEG). It offers temporary spaces in which solutions for technical, economic, and regulatory challenges relating to energy transition can be developed and demonstrated (Federal Ministry for Economic Affairs and Energy, 2019). Moreover, a scheme for regulatory sandboxes has been established to test technical and non-technical innovations in real life and on an industrial scale in critical areas of energy transition. As the smart cities project aims to test various possibilities for digitalization and ensure a good fit with sustainable and integrated urban development, the Federal Ministry of the Interior, Building, and Community has been funding the project since 2019.

For autonomous vehicles, the Federal Ministry of Transport and Digital Infrastructure (BMVI) established the Digital Motorway Test Bed to allow testing of the latest automated driving technology in a real-life setting. The Hamburg Electric Autonomous Transportation project (HEAT) investigates how fully autonomous or self-driving electric minibuses can be safely deployed to transport passengers on urban roads. Since the test vehicles are powered vehicles with highly or fully automated driving functions, the implementation of the project and registration of the cars necessitates applications according to the German Road Vehicles Registration and Licensing Regulations, with exemptions. Regulatory sandboxes can also be designed as testbeds for broad-based participation. The Baden-Württemberg Autonomous Driving Testbed is a regulatory sandbox for mobility concepts that permits companies and research establishments to test technologies and services in the field of connected and automated driving. The combination of various elements of relevance to mobility and the consortium of scientific and municipal partners creates a platform on which key insights and momentum can be gained for the ongoing development of legislation and policy for autonomous driving.

The approach of regulatory sandboxes has been identified as an essential policy instrument for promoting responsible innovation in the national strategy for AI of Norway (Norwegian Ministry of Local Government and Modernisation, 2020). In this strategy, the concept refers to legislative amendments that allow trials within a limited geographical area or period, as well as more comprehensive measures in areas where the relevant supervisory authority needs close monitoring and supervision. The government has established regulatory sandboxes in the field of transportation in the form of legislative amendments that allow testing activities. An act to enable pilot projects on autonomous vehicles came into force in January 2018. Maritime authorities established the first test bed for autonomous vessels in 2016, and two more test beds have been approved since then. In 2019 parliament adopted a new Harbours and Fairways Act, which permits autonomous coastal shipping. Such permission allows sailing in specific fairways, subject to compulsory pilotage or in areas where no pilotage services are provided. Where pilot projects deviate from applicable laws and regulations, they can be conducted with statutory authority in special rules. Alternatively, under the Pilot Schemes in Public Administration Act, public administration can apply to the Ministry of Local Government and Modernisation to deviate from laws and regulations to test new ways of organizing their activities or performing their tasks for a period of up to four years.

In the UK, technology suppliers and their National Health Service (NHS) partners who were delivering machine learning applications in diagnostic pathways have begun work on a regulatory sandbox (Care Quality Commission, 2020). The Care Quality Commission (CQC) formed a team with members from across different functions, as well as a governance committee to oversee the work. The National Institute for Clinical Excellence (NICE), the Medicines and Healthcare Products Regulatory Agency (MHRA), and the NHSX – a joint unit between the NHS and the Department of Health and Social Care to drive the digital transformation of health care – were also included as government partners in this sandbox. They have been working to explore new guidance for NHS providers on AI systems with the

Information Commissioner's Office. The first output from the regulatory sandbox process is a common understanding of what should be present to help deliver high-quality care when using machine learning applications in clinical diagnostics. Developing this shared view of quality with people who use services, providers, technology suppliers, and system partners has been the basis of their work in the sandbox.

In Europe, deregulated special zones have mainly been applied in the fields of self-driving cars and drones. The Swedish government sponsored the world's first large-scale autonomous driving pilot project in 2016. In Belgium, the first special zone for the testing of drones in open labs was established in Antwerp harbor in January 2019. The Russian government has also announced that a new experimental legal framework will be applied to the city of Moscow for AI experimentation.

Given that these various initiatives to create regulatory sandboxes for AI-based innovation have only recently been introduced, it is difficult to make concrete judgments about what impacts have been made by the regulatory sandboxes. There are only limited empirical data from which to draw any conclusions as to the extent regulatory sandboxes have succeeded in creating innovation as expected. At the same time, we do not yet fully comprehend the scope of privacy violations or security risks that consumers may be subjected to by AI algorithms.

## Regulatory Sandboxes for Data Governance in Smart Cities

Although empirical findings are still limited, we can identify a number of key challenges in designing and implementing regulatory sandboxes for AI-based innovation in real-life settings. These include: how to guarantee compliance with regulations for safety, health, environment, security, and privacy, and to what extent regulations can be modified; how to share responsibility between the public and private sectors when accidents or problems have occurred; and how to manage accessibility, sharing, ownership, and use of data. In particular, data governance is a critical

challenge in fully utilizing the approach of regulatory sandboxes for AI-based innovation in the context of smart cities.

Various sectors are undergoing significant transformations by introducing data-driven innovation in smart cities. In the energy sector, distributed energy systems with peer-to-peer exchange of energy have become possible through blockchain technology, with photovoltaics provided through Solar-as-a-Service (SaaS). Smart meters and IoT technologies are providing highly sophisticated services for energy, health, and security to buildings and houses. In transportation, connected, autonomous, sharing, and electrified (CASE) challenges are radically changing the technologies and systems in the sector, and Mobility-as-a-Service (MaaS) is being explored aggressively through alliances among key players across the globe. In the health sector, Software as a Medical Device (SaMD) is being explored, and the diagnosis of cancers based on image recognition is considered especially promising.

An essential approach to stimulating data-driven innovation in smart cities is to foster data collection and sharing. A vast amount of various kinds of data would be collected from energy systems, public transportation, individual vehicles, and buildings, and many benefits would be expected from using that data for different types of innovation. For example, while the data collected through smart meters on energy consumption in households would be useful for optimizing energy use, that data could also be used for providing other services such as home delivery services. The data could tell delivery operators when residents would be at home, allowing them to adjust when to visit the house (Ohsugi & Koshizuka, 2018). The same data could also be used to provide health and security services to the residents of the house.

An open data approach facilitates collaborative efforts among stakeholders to create innovation for smart cities. In comparison to the conventional model of open innovation, which focuses on bilateral collaboration between firms, open innovation 2.0 is a new mode of innovation based on integrated

collaboration through experimentation with a wide range of actors in different sectors (Curley & Salmelin, 2018). Open data initiatives are increasingly considered as defining elements of emerging smart cities, which can be characterized as open innovation economies enabled by the participation of city residents, civic society, software developers, and local small- and medium-sized enterprises (SMEs) (Ojo, Curry & Zeleti, 2015). A recent study which analyzed patent applications in smart cities across the globe suggests that smart city policies have a positive impact on the rate of innovation, particularly in the high-tech sector (Caragliu & Del Bo, 2019).

There are many issues that we need to consider when implementing open data in smart cities. These include the types of data collected, who owns and has access to the data, for what purposes can the data be used, how the data are managed, and what incentives are provided to encourage data sharing to stimulate innovation while addressing concerns about privacy and security in smart cities. Although laboratory-level attempts have been made to integrate various types of datasets and sources on research data scattered across organizations, the scope and amount of data collected and shared needs to be expanded to scale-up innovative initiatives for actual implementation in smart cities. The quality control, error monitoring, and cleaning of data, as well as interoperability between various data standards, must be maintained to secure reliability. Organizational and legal frameworks need to be established concerning the ownership and accessibility of data, and to protect privacy and sensitive data. At the same time, it is also essential to keep a balance between open and proprietary data (Organisation for Economic Co-operation and Development, 2015b).

The collection and use of an extensive range of data, in particular, raises societal concerns in developing smart cities. The case of Sidewalk Toronto – a smart city project initiated in Toronto's waterfront by Alphabet, the parent company of Google – illustrates the seriousness of the concerns among citizens. There are various benefits expected to be provided to the residents and workers in the area, such as

ubiquitous high-speed Internet, intelligent traffic lights, smart shades in public spaces, underground delivery robots, and smart energy grids (Knight, 2019). The smart city plan would generate large quantities of data that could be used to optimize and improve technologies and services. However, some citizen groups were very concerned about the management of the collected data, and the Canadian Civil Liberties Association sued the City of Toronto in an attempt to block the project. After extensive consultation with citizens and companies in the city, the Master Innovation and Development Plan (MIDP) for Toronto was released in June 2019 (Sidewalk Labs, 2019a). The new plan emphasized community engagement and understanding of local needs in response to the concerns raised about building smart cities that are capable of tracking their inhabitants in unprecedented detail. Despite these efforts, the smart city project was eventually terminated (Doctoroff, 2020).

In trying to establish appropriate systems of data governance, it is useful to classify various types of data available in smart cities. Urban data can be defined as including personal, non-personal, aggregate, and de-identified data collected and used in physical or community spaces where meaningful consent before collection and use is difficult to obtain (Sidewalk Labs, 2019b). Non-personal data does not identify an individual and can include other types of non-identifying data not concerning people, such as machine-generated data about weather and temperature, and data on maintenance needs for industrial equipment. Aggregate data is about people in the aggregate and not about a particular individual, and is useful for answering research questions about populations or groups of people. Aggregate counts of people in an office space, for example, can be used in combination with other data, such as weather data, to develop an energy-efficiency program. De-identified data concerns an individual that was identifiable when the data was collected but has subsequently been made non-identifiable. Third-party apps and services can use properly de-identified data for research purposes, such as comparing neighborhood energy usage across a city. Personal data is usually the subject of privacy laws and includes any information

that could be used to identify an individual or that is associated with an identifiable individual. Individuals typically share their personal data with governments and businesses when applying for a license, shopping, or ordering a delivery service.

Digital transparency can be enhanced by providing easy-to-understand language that clearly explains the nature of data and privacy implications of digital technologies to citizens in smart cities (Lu, 2019). Through digital transparency, people are able to understand how and why data is being collected and used in the public realm through a visual language. For example, one hexagon conveys the purpose of the technology; another shows the logo of the entity responsible for the technology; and a third contains a QR code that takes the individual to a digital channel where they can learn more. In situations where identifying information is collected, a privacy-related colored hexagon can also be displayed by combining the technology type (video, image, audio, or otherwise) with the way that identifiable information is used (yellow for identifiable and blue for de-identified before first use, among others). This kind of approach could facilitate citizens' understanding and engagement in smart city projects.

A key question is what would be an appropriate governance system for urban data to maximize the potential of data-driven innovation while minimizing risks to individuals and communities. One approach is to establish a data trust, which is defined as a legal structure that provides for independent stewardship of data (Hardinges, Wells, Blandford, Tennison & Scott, 2019). With data trusts, the organizations that collect and hold data permit an independent institution to make decisions about who has access to data under what conditions, how that data is used and shared and for what purposes, and who can benefit from it. An independent urban data trust would be able to manage urban data and make it publicly accessible by default if appropriately de-identified (Sidewalk Labs, 2019b). An accountable and transparent process for approving the use or collection of urban data would ensure that local companies, entrepreneurs, researchers, and civic organizations can use urban

data. These data would be kept by the data trust and not be sold, used for advertising, or shared without the residents' permission.

In Japan, the Super City Initiative was started in October 2018 in an attempt to respond to the challenge posed by the fourth industrial revolution involving AI and IoT (Secretariat for the Promotion of Regional Development, 2020). The initiative requires that projects go beyond demonstrating a single technology, such as autonomous vehicles in a specific field, and to integrate it with other advanced services, such as cashless transactions and once-only application for administrative procedures, to comprehensively address a societal issue in a city. It also emphasizes that projects should incorporate the views and perspectives of the people living there, not simply the ideas promoted by the developers and suppliers of technologies. The super city initiative provides a particular legal procedure for deregulation that is specifically designed to simultaneously support regulatory reforms in different fields in an integrated manner. The broad regulatory changes involved in building smart cities often require dealing with multiple government agencies. In such cases, a top-down approach is taken; if a municipality obtains approval for smart city plans from its residents, the prime minister in the central government can direct agencies to make exceptions to the relevant regulations as needed. In June 2020, Japan's parliament just passed the "super city" bill, and the government is expected to soon begin taking applications from municipalities, with approvals starting in the summer (Miki, 2020).

In a super city, a data linkage platform plays a crucial role in facilitating close coordination among various services as the operating system (OS) of the city (Secretariat for the Promotion of Regional Development, 2020). A data linkage platform would be developed by professional vendors and operated by local governments, whereas private service providers would offer various services. As long as the residents of the super city agree, it would also be possible for either public agencies or private enterprises to provide services and the platform, making consent by the residents particularly crucial in data governance. For

example, when there are two separate systems for making taxi reservations and doctors' appointments, a data linkage platform can optimize taxi dispatching and appointment scheduling by connecting the relevant data in the two systems. The data linkage platform does not necessarily need to maintain an extensive central database, as data can be stored in separate databases in a distributed way. The providers of digital data and services are required to make their application program interfaces (APIs) open to the public, so that any information system can be developed through the data linkage platform. The super city initiative provides the operator of the data linkage platform with a right to request national and local governments and private enterprises to provide necessary data.

Several issues need to be addressed concerning data governance in smart cities through regulatory sandboxes. For the use of sophisticated services available in smart cities, personal data will be required on various aspects of the residents' lives. In the case of introducing an app connecting taxi–hospital reservations, the data linkage platform would ask the national or local government for personal data on the address, health status, and level of care needed by the elderly. The provision of such data would require the consent of the person in question in accordance with the law. On the other hand, relevant laws might allow the provision of such data without the permission of the person if there is a particular reason, such as contributing to the public interest. As local governments, businesses, or regional councils would make decisions in such cases, clear, transparent, and inclusive procedures are necessary for relevant stakeholders.

Another issue is how to reach a consensus among residents in smart cities. As residents are expected to agree on what kind of city they would like, and which areas they would target, the process of building a consensus needs to be well-integrated into the planning process. Furthermore, the methodologies and procedures for consensus-building need to be specified and institutionalized in an open and inclusive manner. It is also essential to consider how

to protect the rights of those residents who do not want to participate in the data governance scheme of smart cities. Residents need to form a consensus on where the balance should be located between the convenience of the advanced services that rely upon personal data and the risk of the data being used without their consent.

At the same time, the openness and interoperability of data in smart cities needs to be secured. In smart cities, it is often challenging to provide a cross-sectoral service because, typically, data is independent for each field and organization. Reusing and deploying such services to other cities is also difficult because the data system is specialized for each city. Moreover, the cost and labor required for functional expansion in the conventional data system increases, and services cannot easily be expanded to a larger scale. The provision of various services will be improved through close linkage and coordination of data in other systems and cities. APIs play a particularly significant role in facilitating interoperability and data flow. The design process of APIs defines conventions of data exchanges that influence interactions among the stakeholders involved (Raetzsch, Pereira, Vestergaard & Brynskov, 2019). It is essential to make APIs open, secure, and transparent, so that various kinds of data and sophisticated services are connected efficiently and effectively.

Coordinated efforts to share experiences in regulatory sandboxes at the international level will help to foster openness and interoperability to promote data sharing and use for innovation and transparency, as well as trust in managing and governing data to address concerns about privacy and security. So far, no global policy framework has yet been established on how to govern data for smart cities (Russo, 2019). For example, there is no shared set of rules concerning how sensor data collected in public spaces, such as by traffic cameras, should be used. It is of critical importance to explore guidelines and principles for the development and deployment of emerging technologies for smart cities by sharing good practices. As an international initiative to address these challenges, the G20 Global

Smart Cities Alliance on Technology Governance was launched in October 2019. The initiative aims to establish global standards for data collection and use, foster greater transparency and public trust, and promote best practices in smart city governance (World Economic Forum, 2019). Working together with municipal, regional, and national governments, as well as private-sector partners and city residents, the alliance intends to co-design, pilot, and scale-up policy solutions to help cities responsibly implement data-driven innovation. Such an international initiative will contribute to developing a global policy framework for smart cities by examining key issues concerning data governance, including privacy, transparency, openness, and interoperability, based on experiences through regulatory sandboxes in different locations.

## Conclusion

Data-driven innovation plays a crucial role in tackling sustainability challenges. As the development of AI is accelerated by deriving new and significant insights from the vast amount of data generated during the delivery of services every day, training and adaptation is key to creating data-driven innovation. The development of cyber-physical systems such as smart cities is facilitated through the ready availability of and accessibility to data, and its mutual exchange and sharing with stakeholders in different sectors. Hence the new mode of data-driven innovation requires open, dynamic interactions with stakeholders possessing and generating various kinds of data. Close cooperation and collaboration in regards to data is crucial in the innovation process, from the development of novel technologies to deployment through field experimentation and legitimation in society.

It is critical to establish a proper system to govern data-driven innovation in the context of accelerating technological progress and deepening interconnection and interdependence. The speed of technological change with AI is remarkably fast, and it is accompanied by a significant degree of uncertainty in terms of consequences and side effects. Various types of technologies are increasingly becoming

interconnected and interdependent through data exchange and sharing among multiple sectors in smart cities, such as energy, buildings, transportation, and health. These characteristics make it difficult to explain or understand the process of innovation, and contribute to giving rise to a widening gap between technological and institutional changes. AI-based innovation becomes robust by involving the stakeholders who will interact with the technology early in development, obtaining a deep understanding of their needs, expectations, values, and preferences, and testing ideas and prototypes with them throughout the entire process.

Specifically designating geographical areas or sectoral domains, in the form of regulatory sandboxes, can facilitate data-driven innovation by allowing experimental trials of novel technologies and systems that cannot currently operate under the existing regulations. They provide a limited form of regulatory waiver or flexibility for firms to test new products or business models with reduced regulatory requirements, while preserving certain safeguards to ensure appropriate consumer protection. The aim is to provide a symbiotic environment for innovators to test new technologies, and for regulators to understand their implications for industrial innovation and consumer protection. Regulatory sandboxes help to identify and better respond to regulatory breaches by enhancing flexibility and adjustment in regulations, which would be particularly relevant in highly regulated industries, such as the finance, energy, transport, and health sectors.

The approach of regulatory sandboxes will play an especially essential role in governing data-driven innovation in smart cities, which inevitably faces a difficult challenge of collecting, sharing, and using various kinds of data for innovation while addressing societal concerns about privacy and security. Regulatory sandboxes can relax or adjust some of the relevant regulations, so that these new technologies can be tested for actual adoption and use. How regulatory sandboxes are designed and implemented can be locally adjusted, based on the specificities of the economic and social conditions and contexts,

to maximize the effect of learning through trial and error. Various types of new promising technologies can be verified, adopted, and integrated, effectively improving technological performance, reliability, and integration, as well as contributing to cost reduction. As AI-based innovation involves rapid technological change, uncertain market developments, and diverse social norms, there are many economic, ethical, and legal issues comprised of various interests and preferences. Regulatory sandboxes need to be flexible to accommodate the uncertainties of innovation, and precise enough to impose society's preferences on emerging innovation, functioning as a nexus of top-down strategic planning and bottom-up entrepreneurial initiatives.

Emerging cases of regulatory sandboxes for smart cities show that data governance is critical to maximizing the potential of data-driven innovation while minimizing risks to individuals and communities. With data trusts, the organizations that collect and hold data permit an independent institution to make decisions about who has access to data under what conditions, how that data is used and shared and for what purposes, and who can benefit from it. Alternatively, a data linkage platform can facilitate close coordination between the various services provided and the data stored in a distributed manner, without maintaining an extensive central database. The operator of the data linkage platform would require a right to request national and local governments and private enterprises to provide necessary data. APIs-linking data and services need to be open to the public so that any information system can be developed through the data linkage platform.

It is critically important that the data governance systems of smart cities are open, transparent, and inclusive. While the provision of personal data would require the consent of the person in question, the relevant law might allow the provision of such data without the permission of the person if there is a particular reason, such as contributing to the public interest. As local governments, businesses, or regional councils would be expected to make a decision, clear, transparent, and inclusive procedures are necessary

for relevant stakeholders. The process of building a consensus among residents needs to be well-integrated into the planning of smart cities, with the methodologies and procedures for consensus-building specified and institutionalized in an open and inclusive manner. It is also essential to respect the rights of those residents who do not want to participate in the data governance scheme of smart cities. As APIs play a crucial role in facilitating interoperability and data flow in smart cities, open APIs will facilitate the efficient connection of various kinds of data and sophisticated services. International cooperation will be critically important to develop common policy frameworks and guidelines for facilitating open data flow while maintaining public trust among smart cities across the globe.

## Policy Recommendations

**Recommendation 1:** New policy approaches are required to govern data-driven innovation in the context of accelerating technological progress and deepening interconnection and interdependence.

**Recommendation 2:** Regulatory sandboxes should be established to facilitate data-driven innovation by allowing experimental trials of novel technologies and systems that cannot currently operate under the existing regulations through specifically designating geographical areas or sectoral domains.

**Recommendation 3:**. Stakeholders should be involved from the early stages of technological development in order to obtain a deep understanding of their needs, expectations, values, and preferences, and to test ideas and prototypes with them throughout the entire process.

**Recommendation 4:** Regulatory sandboxes should be designed and implemented by incorporating the specificities of local economic and social conditions and contexts to maximize the effect of learning through trial and error.

**Recommendation 5:** Regulatory sandboxes need to be flexible to accommodate the uncertainties of innovation, and precise enough to impose society's preferences on emerging innovation, functioning as a nexus of top-down strategic planning and bottom-up entrepreneurial initiatives.

**Recommendation 6:** Data governance systems of smart cities should be open, transparent, and inclusive to facilitate data sharing and integration for data-driven innovation while addressing societal concerns about security and privacy.

**Recommendation 7:** The procedures for obtaining consent on the collection and management of personal data should be clear and transparent to relevant stakeholders with specific conditions for the use of such data for public purposes.

**Recommendation 8:** The process of building a consensus among residents should be well-integrated into the planning of smart cities, with the methodologies and procedures for consensus-building specified and institutionalized in an open and inclusive manner.

**Recommendation 9:** Application programming interfaces (APIs) should be open to facilitate interoperability and data flow for efficient connection of various kinds of data and sophisticated services in smart cities.

**Recommendation 10:** Common policy frameworks should be explored to develop guidelines for data collection and use, foster greater transparency and public trust, and promote interoperability and open data flow among smart cities across the globe.
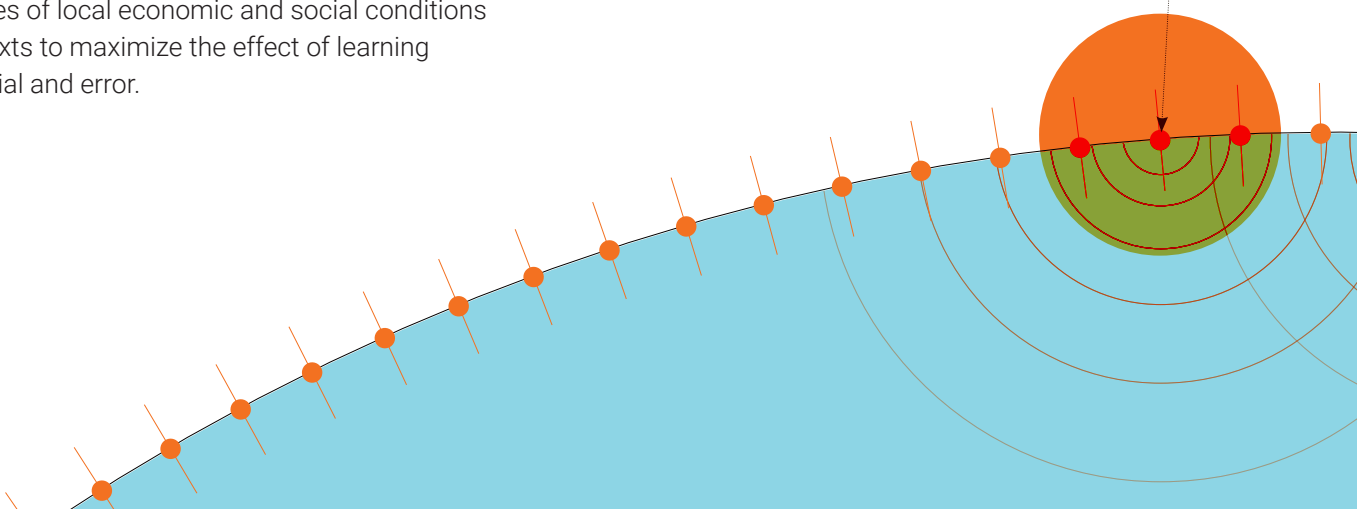
# References

Allen, H. J. (2019). Regulatory Sandboxes. *George Washington Law Review, 87*(3), 579-645.

Allen, H. J. (2020). Sandbox Boundaries. *Vanderbilt Journal of Entertainment & Technology Law, Forthcoming, 22*(2), 299-321.

Beede, E. (2020, April 25). *Healthcare AI systems that put people at the center.* Retrieved from Google Blog: https://www.blog.google/technology/health/healthcare-ai-systems-put-people-center/

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020). A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-12.

Boyd, A. (2018, May 9). *10 Drone Programs Get Federal OK To Break The Rules.* Retrieved from Nextgov: https://www.nextgov.com/emerging-tech/2018/05/10-drone-programs-get-federal-ok-break-rules/148098/

Caragliu, A., & Bo, C. F. (2019). Smart innovative cities: The impact of Smart City policies on urban innovation. *Technological Forecasting and Social Change*, 142, 373-383.

Care Quality Commission. (2020, March). Using machine learning in diagnostic services: A report with recommendations from CQC's regulatory sandbox. *Care Quality Commission*.

Curley, M. (2016). Twelve principles for open innovation 2.0. *Nature*.

Curley, M., & Salmelin, B. (2018). Data-Driven Innovation. In Open Innovation 2.0: *The New Mode of Digital Innovation for Prosperity and Sustainability*. Cham: Springer International Publishing.

Doctoroff, D. L. (2020, May 7). *Why we're no longer pursuing the Quayside project — and what's next for Sidewalk Labs*. Retrieved from Medium: https://medium.com/sidewalk-talk/why-were-no-longer-pursuing-the-quayside-project-and-what-s-next-for-sidewalk-labs-9a61de3fee3a

Energy Market Authority. (2017, October 23). Launch of Regulatory Sandbox to Encourage Energy Sector Innovations. *EMA*.
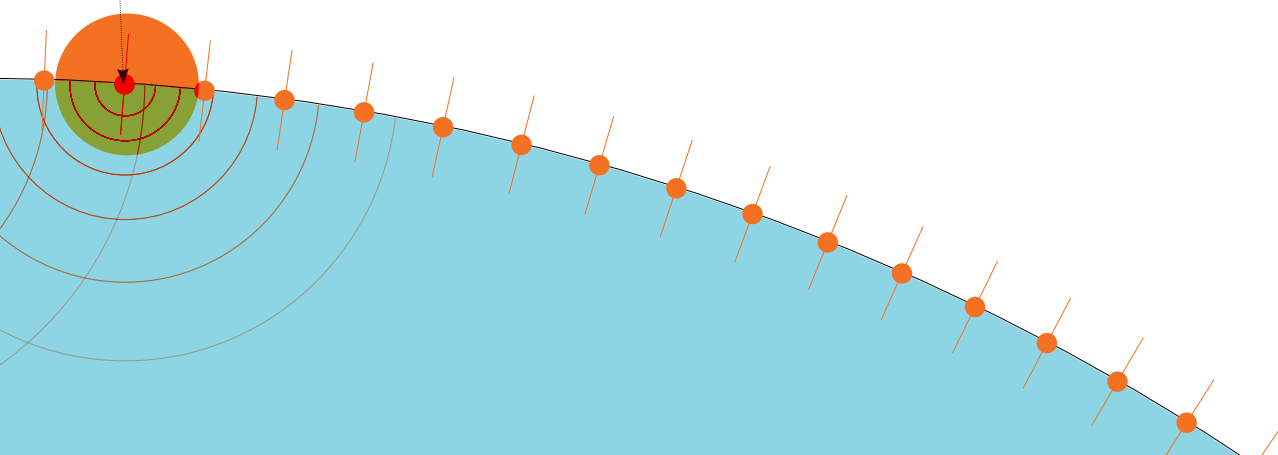
Federal Aviation Administration. (2019, December 10). *UAS Integration Pilot Program*. Retrieved from United States Department of Transportation: **https://www.faa.gov/uas/ programs_partnerships/integration_pilot_program/**

Federal Ministry for Economic Affairs and Energy. (2019, July). Making Space for Innovation: The handbook for regulatory sandboxes. *BMWi*.

Financial Conduct Authority. (2015). Regulatory Sandbox. *FCA*.

Financial Conduct Authority. (2017, October). Regulatory Sandbox Lessons Learned Report. *FCA*.

Fleischer, V. (2010). Regulatory Arbitrage. *Texas Law Review, 89*(2), 227-289.

Food and Drug Administration. (2019, April). Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. *FDA*.

Guston, D. H. (2014). Understanding 'anticipatory governance'. S*ocial Studies of Science, 44*(2), 218-242.

Hardinges, J., Wells, P., Blandford, A., Tennison, J., & Scott, A. (2019, April). Data trusts: lessons from three pilots. *Open Data Institute*.

Iizuka, M., & Ikeda, Y. (2019). "Regulation and innovation under Industry 4.0: Case of medical/healthcare robot, HAL by Cyberdyne." Working Paper Series #2019-038, Maastricht Economic and Social Research Institute on Innovation and Technology (UNU-MERIT).

Ikeda,Y., & Iizuka, M. (2019, October). "International Rule Strategies for Implementing Innovation in Society: A Case Study of the Medical Healthcare Robot HAL." RIETI Policy Discussion Paper Series 19-P-016, Research Institute of Economy, Trade and Industry.

Information Commissioner's Office. (2019). ICO opens Sandbox beta phase to enhance data protection and support innovation. *ICO*.

Knight, W. (2019). Alphabet's smart city will track citizens, but promises to protect their data. *MIT Technology*.

Lu, J. (2019, April 19). *How can we bring transparency to urban tech? These icons are a first step*. Retrieved from Medium: **https://medium.com/sidewalk-talk/how-can-we-make-urban- tech-transparent-these-icons-are-a-first-step-f03f237f8ff0**

Miki, R. (2020, May 13). *Coronavirus pushes Japan closer to high-tech 'super cities'*. Retrieved from Nikkei Asian Review: **https://asia.nikkei.com/Politics/Coronavirus-pushes-Japan-closer-to-high-tech-super-cities**

Norwegian Ministry of Local Government and Modernisation. (2020). National Strategy for Artificial Intelligence. *H-2458 EN*.

OECD. (2015a). Data-Driven Innovation: Big Data for Growth and Well-Being. *OECD Publishing*.

OECD. (2015b). Making Open Science a Reality. *OECD*.

OECD. (2018). OECD Science, Technology and Innovation Outlook 2018.

OECD. (2019). Digital Innovation: Seizing Policy Opportunities. *OECD*.

Office of Gas and Electricity Markets. (2018a). Enabling trials through the regulatory sandbox. *ofgem*.

Office of Gas and Electricity Markets. (2018b). Insights from running the regulatory sandbox. *ofgem*.

Ohsugi, S., & Koshizuka, N. (2018). Delivery Route Optimization Through Occupancy Prediction from Electricity Usage. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, 842-849.

Ojo, A., Curry, E., & Sanaz-Ahmadi, F. (2015). A Tale of Open Data Innovations in Five Smart Cities. *2015 48th Annual Hawaii International Conference on System Sciences (HICSS-48)*, 2326-2335.

Pagallo, U., Casanovas, P., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., . . . Valcke, P. (2019). On Good AI Governance: 14 Priority Actions, A S.M.A.R.T. Model of Governance, and a Regulatory Toolbox. *AI4People*.

Pollman, E. (2019). Tech, Regulatory Arbitrage, and Limits. *European Business Organization Law Review, 20*(3), 567-590.

Raetzsch, C., Pereira, G., Vestergaard, L. S., & Brynskov, M. (2019). Weaving seams with data: Conceptualizing City APIs as elements of infrastructures. *SAGE Journals, 6*(1).

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., . . . Larochel. (2019). Machine behaviour. *Nature, 568*(7753), 477-486.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision, 115*(3), 211-252.

Russo, A. (2019). World Economic Forum to Lead G20 Smart Cities Alliance on Technology Governance. *World Economic Forum*.

Secretariat for the Promotion of Regional Development. (2019). *The National Strategic Special Zones*. Retrieved from Cabinet Office, Prime Minister's Office of Japan: **https://www.kantei.go.jp/jp/singi/tiiki/kokusentoc/supercity/supercityforum2019/ supercityforum2019_EnglishVer.html**

Secretariat for the Promotion of Regional Development. (2020). About the Super City Initiative. *Cabinet Office, Prime Minister's Office of Japan*.

Sidewalk Labs. (2019a). Sidewalk Labs Publishes Comprehensive Blueprint for the Neighbourhood of the Future. *Sidewalk Labs*.

Sidewalk Labs. (2019b). Toronto Tomorrow: A new approach for inclusive growth, Volume 2. *Sidewalk Labs*.

Stilgoea, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy, 42*(9), 1568-1580.

Taeihagh, A., & Lim, H. S. (2019). Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Reviews, 39*(1), 103-128.

World Economic Forum. (2019). Forum-led G20 Smart Cities Alliance will create the first global framework for smart city governance. *World Economic Forum*.

Yarime, M. (2017). Facilitating data-intensive approaches to innovation for sustainability: opportunities and challenges in building smart cities. *Sustainability Science, 12*(6), 881-885.

Zetzsche, D. A., Buckley, R. P., Arner, D. W., & Barberis, J. N. (2017). Regulating a Revolution: From Regulatory Sandboxes to Smart Regulation. *Fordham Journal of Corporate and Financial Law, 23*(1), 31-103.

# How to Expand The Capacity of AI to Build Better Society

# Including Women in AI-enabled Smart Cities: Developing Gender-inclusive AI Policy and Practice in the Asia-Pacific Region
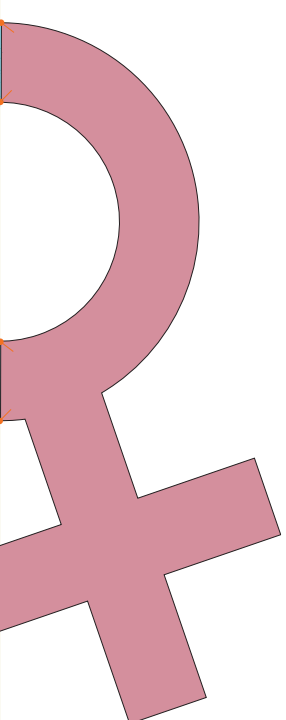
**Caitlin Bentley***
Lecturer in
AI-enabled Information Systems,
Information School,
University of Sheffield
Honorary Fellow
3A Institute,
Australian National University

**Katrina Ashton,
Brenda Martin,
Elizabeth Williams,
Ellen O'Brien,
Alex Zafiroglu, and
Katherine Daniell**
3A Institute,
Australian National University

# 1. Introduction

Heralded as the answer to rapid urbanization and related environmental, social, and governance challenges, smart city developments are proliferating across the Asia-Pacific region. Sensors, artificial intelligence (AI), machine learning algorithms, actuators, and other advanced technologies are being built into city infrastructures. AI-enabled systems undertake advanced data analytics, feeding into predictions and automated decision-making that are enacted through actuators or other system structures. These new AI-enabled systems are designed to tackle pressing urban issues such as air pollution, traffic congestion, and public safety.[1]

However, not everyone has benefited from smart city developments. For instance, Cathelat (2019) demonstrates that a gender dimension is lacking within smart city plans, even when there is an expressed commitment to social inclusion. Broadly, women are also less connected to the Internet, and technology access inequalities by gender are on the rise in Asia-Pacific (Sey & Hafkin, 2019). AI-enabled systems introduce new risks and security concerns that may disproportionately affect women (Finlay, 2019). It is important to understand and promote effective ways to design, develop, manage, and regulate AI-enabled systems more inclusively with and for women.

AI-enabled systems affect multiple aspects of women's lives, as computational modelling increasingly informs numerous areas of urban governance. Women may interact with AI-enabled smart cities through multiple touchpoints, including embedded sensors, Internet and mobile networks, and other networks (workplaces, healthcare, transportation, retail centers, etc.). Ultimately, data streams record women's behaviors, preferences, locations, and values. Data streams may then be analyzed and incorporated into machine learning algorithms, through which specific predictions are made. Due to the capacity for real-time analytics, as well as the dominant focus of these systems on prediction and prevention, AI-enabled smart cities suggest the need for an approach that is cognizant of the social dynamics at play and of the cultural richness and diversity of our communities.

This work has two interrelated goals: to include the voices, theories, experiences, and histories of female and feminist scholars and activists in developing better policies for AI-enabled smart cities; and to evaluate a practical and concrete framework that policymakers can use to support women, while taking into account the specific opportunities and risks introduced by AI in smart city initiatives. Towards these ends, this paper critically reviews the extant literature, focusing specifically on the status of AI-enabled smart city initiatives across multiple countries in the Asia-Pacific region. We then analyze two key applications

---

1. ASEAN (2018) Smart City Network progress report gives a good overview of the 26 pilot initiatives underway across eight countries.

of AI for social good used within smart city initiatives: public safety and transportation. In general, we find limited evidence of gender-responsive policymaking and practice, and little empirical research concerning how AI contributes to safer public spaces or more effective transportation systems for women. We argue that greater integration between the technical capacity of AI-enabled systems and diverse communities of women is needed.

We introduce and evaluate the 3A Framework as an effective approach to leading and forming such integration holistically. Policymakers need practical and concrete ways to support women, whilst taking into account the specific technological shifts underway due to AI. This Framework provides a set of core questions that can be used as a starting point. This research maps out key insights generated from interviews with leading female and feminist scholars and activists who have significant knowledge and experience of working in Asia-Pacific. The experts reflected on what inclusive practice means when it comes to working at the intersection of gender and advanced technologies. We examine how these insights can be used to elaborate on the Framework, thereby establishing a method for inclusive policymaking and practice.

## 2. Smart cities in the Asia-Pacific region: are they inclusive to women?

Hojer and Wangel (2015) argue that the idea of a smart city has its roots in concepts of "cybernetically planned cities" developed in the 1960s, in which networked and computational capabilities would be built into urban development plans starting in the 1980s, mostly within the US and Europe. The concept has raised significant worldwide debate due to the tensions in its instrumental meaning versus associated intended outcomes (Allwinkle & Cruickshank, 2011; Hollands, 2008; Kitchin, 2014). AI-enabled smart cities are increasingly common and are tied to the spread of connected Internet of Things (IoT) devices and advances in computing power. That said, such systems are envisaged, implemented, and regulated in diverse ways across the varied social, political, and economic landscapes of the Asia-Pacific region.

Whether Asia-Pacific smart cities are inclusive to women depends on what we mean by being "inclusive

to women", and on the management model that is implemented. Concerning being "inclusive to women", we adopt the UN DESA (n.d.) definition:

> *Social inclusion is the process by which efforts are made to ensure equal opportunities that everyone, regardless of their background, can achieve their full potential in life. Such efforts include policies and actions that promote equal access to (public) services, as well as enable citizens' participation in the decision-making processes that affect their lives.*

This definition considers the inclusion of women as a societal issue, falling under the remit of multiple actors and institutions. It does not mean that women's issues and perspectives are favored over men's, rather, it says that women's access to services and decision-making processes need to be considered in context and in relation to others. However, we acknowledge that the UN definition may privilege notions of "equality" and "access" over "equity" and "outcomes." Roces (2010) details how international feminist discourses have conflicted and resonated in different ways with various Asian feminist movements. Our research examines the 3A Framework as a means for policymakers and practitioners across a range of Asia-Pacific cultures to generate context specific goals, definitions, and outcomes of gender inclusiveness, and to better understand how they play out in smart city contexts.

Across the Asia-Pacific region, many countries are organizing state-level smart city initiatives, with many making provisions for social inclusion within them (Table 1). We find that there are often no principles or programs identified within these high-level initiatives defining or standardizing how social inclusion should be implemented. Similarly, many Asia-Pacific countries have published national AI strategies, such as India's National Strategy for AI (NITI Aayog, 2018) or Thailand's Industry 4.0 Policy (Baxter, 2017). Countries have also enacted data privacy and protection laws, though it is not clear how all of these policy areas are mediated. For instance, the Australian Human Rights Commission (2019) identified gaps in current law, application of law, regulatory measures, and education and training when evaluating the adequacy of existing laws in protecting human rights in the context of AI. The lack of clarity surrounding national law and policy for AI-enabled smart cities across the region cast doubts over whether the inclusion of women has been a priority, and raises questions regarding the bases of inclusion.

**Cross-country comparison of national smart city plans and social inclusion provisions**

| Country/ Region (ordered by GDP) | National Smart City Policy and Plans | Social Inclusion Notes |
|---|---|---|
| China | China incorporated their Smart City Initiative into its national policy, and is a main reason for accelerated development of over 500 smart city pilots in China (Long, Zhang, Zhang, Chen, & Chen, 2019). | Chan and Anderson (2015) reported a transition in China from technology-centered to human-centered smart cities with a focus on increasing public participation in the country. No details about gender inclusion in national policy were found. |
| Japan | Launch of a "super-smart city" initiative called Society 5.0 in 2016. National framework outlining how AI, IoT devices, and robots will transition Japan from an information society to an AI-enabled society, bringing about a human-centered society (Cabinet Office, Government of Japan, n.d.). | Society 5.0 plans explicitly mention social inclusion goals, focusing on optimal and tailored services for individuals, whilst overcoming national challenges such as the ageing population, social polarization, and depopulation (UNESCO, 2019). |
| India | In 2015, the Indian Government pledged to create 100 smart cities by 2020. Only a portion of allocated funds have been used so far, and the timeframe has been extended to 2023. The Smart Cities Mission is the responsibility of the Ministry of Housing and Urban Affairs. | The Smart Cities Mission Statement and Guidelines (2015, p.6) includes 10 core infrastructure elements, one of which is the "safety and security of citizens, particularly women, children, and the elderly". This is the only context in which gender issues are specifically mentioned. Another document provides examples of citizen engagement activities, including ways to be inclusive (e.g., placing Wi-Fi hotspots in slums) (Government of India – Ministry of Urban Development, 2015). |
| South Korea | In 2013, the federal government launched an initiative to construct ubiquitous cities, which has transitioned through two additional phases to connect and decentralize smart city development across the nation. Since 2018, national policy incorporates testbeds, living labs, and implementation of AI technology (Ministry of Information and Communications, 2019b). | Explicit aim to make South Korean's citizens' lives happy and inclusive in smart cities. A five-year, mid-to-long-term roadmap was established, incorporating this vision into its plans (Ministry of Information and Communications, 2019b). No details in reference to women or gender inclusion were found. |

**Table 1: Cross-country comparison of national smart city plans and social inclusion provisions**

| | | |
|---|---|---|
| **Australia** | The Australian Smart City Plan was released by the Department of Prime Minister and Cabinet in 2016 and now sits with the Department of Transport, Infrastructure, Regional Development, and Communications (Department of the Prime Minister and Cabinet, 2016). | No mention of social inclusion, gender, or citizen participation. |
| **Indonesia** | In 2017, the national government created the "100 Smart Cities Movement" initiated by the Ministry of Communication and Information Technology of the Republic of Indonesia, and in 2018 focused on improving public services and increasing regional competitiveness (Davy, 2019). | Equality is mentioned in a press release on the first phase of their smart cities program (Ministry of Communication and Information – Public Relations Bureau, 2017). Individual city master plans (Laksmi, 2018) include some specific mentions of preventing violence against women and children. Sustainability, when mentioned, includes a focus on social dimensions. Citizen participation is an important part of city and district planning. |
| **Thailand** | Smart City Thailand (2018) is a national program designed to roll out smart city services to all 76 provinces and Bangkok by 2022. It incorporates multiple government divisions, is managed by a dedicated unit called the Digital Economy Promotion Agency (DEPA), and involves multiple private sector actors. | One of the seven dimensions of Thailand's plan is centered on building "Smart People" by improving knowledge and skills of residents in order to "decrease social and economic inequality and provide new opportunities for creativity, innovation, and public participation" (Smart City Thailand, 2018, p. 9). No details were given in reference to women or gender differences specifically. |
| **Hong Kong** | Hong Kong's Office of the Government Chief Information Officer has a Smart City Blueprint focused on embracing technology towards strengthening the economy and achieving a high quality of life (Innovation and Technology Bureau, 2017). | The Blueprint focuses on application areas and does not mention specifics in relation to women. One goal is to nurture young talent to gain skills in science, technology, engineering, and mathematics (STEM), but there is no discussion of gender differences. |

**(Cont.) Table 1: Cross-country comparison of national smart city plans and social inclusion provisions**

| Malaysia | Malaysia's Ministry of Housing and Local Government (2018) released a national framework, outlining its definition, key smart city challenges, national policy, strategic areas of application, indicators, governance arrangements, and pilot project descriptions. | A main criterion given is gender empowerment and inclusivity of vulnerable groups. The seventh of 16 city policies given is "Social inclusion, especially gender equality shall be given emphasis in smart city development" (Ministry of Housing and Local Government, 2018, p. 36). This includes supportive physical infrastructure and programs, as well as participation in decision-making. |
|---|---|---|
| Singapore | Smart Nation Singapore (2020) outlined three pillars of action surrounding the digital economy, digital government, and digital society led by the Smart Nation and Digital Government Office and the Infocomm Media Development Authority. | The digital society blueprint refers to inclusion in terms of digital inclusion but does not refer to the specific needs of women. The digital government likewise tracks citizen satisfaction with its services, but does not explain how they address gender differences, if at all (Smart Nation Singapore, 2018). |
| Vietnam | Ministries and agencies are currently researching and completing building guidelines, mechanisms, and policies for smart cities (Ministry of Information and Communications, n.d.), with a first project launched in October 2019 focusing on air and water quality monitoring, renewable energy, public transport, and others, taking part in the ASEAN network (ASEAN, 2018; Ministry of Information and Communications, 2019a). | No mention of social inclusion, gender or citizen participation. |
| Samoa | Samoa's National Urban Policy (2013) is a good example of how Pacific Islands may instead prioritize issues of sustainability, resilience, and inclusion over technologically-centered smart cities. | Whilst inclusivity is a core mission statement, the Policy does not elaborate on what this means. |

**(Cont.) Table 1: Cross-country comparison of national smart city plans and social inclusion provisions**

The conflicted national-level policy space means that social inclusion is often implemented at the initiative level within a particular city or for a specific purpose. A range of what can be classified as "top-down" and "bottom-up" smart city models can be observed. We refer to a "top-down" approach as one where the locus of control over the design, governance, sensing, computation, and/or acting in an AI-enabled system is centralized in some way. For instance, data streams from various sensors are aggregated onto a "dashboard". Taweesaengsakulthai et al. (2019) compared the top-down projects led by the central government in Nakhon Nayok, Phuket, and Chiang Mai provinces with the more locally-driven, bottom-up approach of the Khon Kaen smart city initiative. They noted that the smart cities in Phuket and Chiang Mai put a strong emphasis on supporting the tourism industry rather than their citizens, and speculated that the reason these provinces were chosen by the central government for the smart city initiative was because they are both highly attractive tourist destinations.

Nevertheless, top-down approaches may facilitate widespread integration and use of computational resources across a network when centralized in some way. For example, where environmental sustainability is concerned, centralized aggregation is being explored for monitoring emissions flows and making continuous adaptations to optimize these emissions (Giest, 2017). This sort of aggregated analysis and anticipatory policymaking may not work well for the inclusion of women because there are fewer known "levers" that enable decision-makers to determine exactly how to respond to certain issues. Some countries are therefore implementing public participation processes to facilitate deliberative decision-making in smart city systems (Chan and Anderson, 2015). However, the examined approaches have not yet addressed how such processes may need to change in the context of AI, nor how unequal power relations between men, women, and LGBTQI+ are addressed.

Alternatively, "bottom-up" approaches typically center on placing participation and accountability towards marginalized people, including women, at their core. Bottom-up approaches are characterized by participatory processes, highlighting how local citizens may know best how to respond to the issues they are confronting in their local area, as with Sadoway and Shekhar's (2014) examination of Transparent Chennai's community-driven approach to smart city governance. In contrast, Trencher (2019) analyses another "bottom-up" smart city initiative in Aizuwakamatsu, Japan, noting that the high level of citizen participation was driven by skilled corporate professionals. Bottom-up approaches may also fail to take into account large sets of interdependent factors, as well as the plural intents, interests, and power relations of the people involved. Moreover, AI could be used to scale applications and services that have wide benefit potential to complement grassroots engagement. A clear national strategy that embraces the benefits and minimizes the drawbacks of both top-down and bottom-up approaches could enable better outcomes for women.

Another complicating factor for women's inclusion is the breadth and diversity of actors involved in planning and managing smart city initiatives. Public-private partnerships (PPPs) are common in Asia-Pacific, with examples in India (SCC India Staff, 2018), Thailand (Huawei Enterprise, 2019), China, South Korea, and Japan (Thrive, 2018). Large technology companies are increasingly expanding their roles from suppliers to smart city co-investors, designers, and managers (Cathelat, 2019). Lam and Yang (2020) examine why PPPs occur, specifically in Hong Kong. They find that in the public sector, the most important criteria were availability of needed data, availability of expertise, possibility to maintain transparency of procurement, and monitoring of operations. In the private sector, the most important criteria were possibility to maintain transparency of procurement and monitoring of operation, complexity of coordinating government

departments, and availability of expertise. It is vital to note the lack of mention or consideration of community relations within this study. It appears that whilst PPPs are crucial for the acquisition of resources and expertise, private sector actors may not hold any responsibility towards citizens.

As a result, many countries are pursuing complementary approaches to address social inclusion concerns. For example, Pune in India developed their own framework to engage their citizens as part of their smart city initiative with mixed success (Ministry of Housing and Urban Affairs, 2015). In contrast, Marsal-Llacuna (2015) and Panori et al. (2019) discuss indicators and multi-dimensional poverty indexes, respectively, as a means to foster socially inclusive outcomes. In other fields, it is well established that participatory citizen engagement processes can help to meet social inclusion objectives, but often only if they are negotiated into the design and implementation in a manner cognizant of these objectives; otherwise, citizen engagement processes can perpetuate existing power-structures, inequalities, and exclusion of certain participant groups (Musadat, 2019; Daniell, 2012). Thus, there is still a need to understand how to design such engagement processes in a way that women's perspectives will not remain marginalized and so that they have the opportunity to influence AI-enabled smart city development.

## 3. AI for social good? Opportunities and risks of AI smart city technology for women

An increasing number of AI-enabled smart city initiatives are aiming to improve the well-being and quality of life of residents and visitors. We are particularly interested in applications that hold significant opportunity and risk for women. Based on our cross-country review of smart city progress in Asia-Pacific, we selected two key smart city applications that have seen substantial AI implementation. The following sections unpack the AI components of these two key applications: public safety and transportation.

### 3.1. Improving the safety and security of women in public spaces through facial recognition technology

Public safety and security issues differ greatly across Asia-Pacific urban contexts. However, there are some safety and security issues that affect certain genders disproportionately (Heise et al., 2002; Jackman, 2006). Multiple accounts across the region reflect the risks and fear that women experience due to sexual harassment, assault, and violence in public spaces (Baruah, 2020; Plan International, 2016; Rao, 2017; UN Women, 2017). In Indonesia, women are 13 times more likely to be harassed in public places than men (Widadio, 2019). Whereas in Mumbai, India, Bharucha and Khatri (2018) found 30% of the women surveyed had been groped in public. Human trafficking and forced labor are two other significant safety and security issues affecting women in the Asia-Pacific region (Global Slavery Index, 2019; World Vision Australia, 2007). These issues also affect men, but trafficking for sexual exploitation makes up a large proportion of human trafficking, and in these cases women and girls are usually the victims (Lee, 2005; Piper 2005). An increasingly common strategy to reduce levels of violence and support intra-regional efforts to curb human trafficking and forced labor is to embed automated facial recognition technology (AFRT) into smart city initiatives. For example, in 2019, thanks to AFRT, India celebrated the matching of 10,561 missing children with those living in institutions (Zaugg, 2019).

However, AFRT is not a silver bullet, having generated significant public debate around the technical limitations of the underpinning AI technology and its implications for individual privacy and centralization of power in urban governance. Public opinions on these matters are nuanced across the region; yet, in all contexts, better informed decisions can be made by understanding the components of this AI system. There are two types of facial recognition systems: verification and identification (Grother et al., 2019). Verification seeks to determine if two images of a face match, whereas identification matches a face shown in an image with potential matches in a database of

images. The main way that AFRT assists in reducing violence in public places is its ability to identify assailants post hoc. Similarly, human trafficking and forced labor also depends on authorities having records of victims and being able to match or identify victims. However, there are still concerns about how well this technology works for different demographics, as well as possible side effects and how effective systems incorporating AFRT are at solving the problems they seek to address. For instance, the National Institute of Standards in Technology (NIST) found that women were significantly more likely to be misidentified than men, with false positive rates two to five times higher (Grother et al., 2019). We outline the potential reasons for misidentification in Appendix 1.

When considering the needs and perspectives of women, there are still many substantial gaps in the knowledge. For instance, it is not clear whether post-hoc identification of perpetrators actually has any bearing on the safety and security of women. New AI applications to detect unusual behavior, rather than matching of perpetrators post hoc, may be beneficial in that regard (see Huawei Enterprise, 2019). However, these applications are in the early stages of development and there is no evidence to support their effectiveness (Barrett et al., 2019). It is also not clear what happens to women once they are identified as victims of human trafficking or forced labor, and whether there are other applications of AI technology to identify trafficking patterns (such as one solution discussed in Section 5.1). Evidence outlining the effectiveness of such systems on crime reduction in the Asia-Pacific region is also lacking.

Lastly, little attention has been paid to data security issues, which may also impinge on the safety and security of women when misuse of the system or data breaches occur. There is also the question of how the information will be used: does an alert go to a human, or will there be an automated intervention? Very little discussion has taken place regarding how the images are stored and for how long, which becomes a significant issue when data is centralized (security

risks) and/or used for multiple purposes (various issues around consent and biometric data ownership). There is a need to consider how AFRT will contribute to socially good outcomes for women by examining AI as part of a wider smart city system.

## 3.2. Increasing mobility for women through AI-enabled transportation systems

Traffic congestion and mobility is a significant challenge in the rapidly growing cities of Asia-Pacific, and is commonly found on the wish list of problems to address within smart city initiatives. It is an issue that impacts on citizen well-being, leading to long hours of commuting, increased air pollution, and inaccessibility of city services. Two of the AI-enabled responses often deployed in smart cities are smart traffic lights and smart public transport.

Smart transportation systems often require major infrastructure works and the development of one or multiple systems to manage and optimize transport at various levels of scale and complexity (see Appendix 1 for a breakdown of these). Much empirical research and development in this field focuses on optimizing traffic flows based on real-time monitoring of traffic conditions (Javaid, 2018; Ghazal et al., 2016; Zhao et al., 2012). Data on traffic conditions is collected using vehicle detection sensors and either used to determine optimal timing for a single traffic light, or transmitted over the Internet to a data processing center where it is automatically analyzed to determine optimal traffic lights for a broader system. Efficiency gains in smart public transport are envisaged in a similar manner (Hörold et al., 2015; Haque et al., 2013). Public transportation services can be integrated within the same traffic management system to both prioritize public transport vehicles over private vehicles at intersections, as well as to inform route optimization to service popular routes effectively and avoid congestion. As such, smart traffic management systems usually include an end-to-end platform to which users have access (usually a mobile application).

It is often not clear how system engineers have encoded priorities into the optimization of transport systems. Women may have particular mobility patterns and concerns that have not been factored into optimization algorithms. Data collection in smart city initiatives is often aggregated across genders, which renders women's specific patterns and needs invisible. Inequities persist in Asian cities despite longstanding evidence of gendered differences in transport and several initiatives to address issues (Thynell, 2016). According to Singh (2019), women often make more complex multi-purpose trips using different modes of transport, travelling at off-peak hours. Women also place a higher priority on safety and security in their transport than men, and this can lead them to take more costly or less efficient modes of transport (Gekoski et al., 2017). Little to no attention has been paid to understand how and why women's mobility can be supported and affected by AI-enabled systems. As such, it is often assumed that smart traffic lights and AI-enabled public transport will serve the interests of women because of efficiency gains in transportation systems. Rather, this needs to be tested and women's preferences on system objectives factored into optimization functions or data sets for learning algorithms – even if these need to initially be synthesized for training purposes.

There has been work done on issues women find important using AI techniques, such as how to make public transport safer. In Australia, Transport New South Wales (2020) recently proposed a challenge to seek tenders for solutions to make travelling in Sydney safer for women at night, with a focus on data and suggested solution areas including "Deep Technology." There have also been non-technical solutions proposed, such as women-only carriages of subways, although some argue that such solutions do not address the root of the problem and are instead reinforcing divisions between the sexes (Thynell, 2016). More work is needed to better integrate the needs and aspirations of women in AI-fueled transportation systems.

## 4. Addressing women's needs and aspirations in AI-enabled smart cities

Overall, we find a lack of clarity in national smart city policymaking concerning the presence and inclusion of women. Our review of two AI for social good applications likewise finds significant gaps in the knowledge concerning how these technologies contribute to making public spaces safer and transportation systems more effective for diverse women. To address the needs and aspirations of diverse women, our approach synthesizes principles, practices, and concerns of female and feminist scholars, activists, and practitioners with significant expertise in supporting women. We sought to interview scholars with experience working at the intersection of women and technology, but also included feminists with broader ranging experience across Asia-Pacific contexts. We conducted 12 interviews with 13 selected scholars, activists, and practitioners (Table 2), and contacted another 23 experts, but were either unable to schedule an interview, had no response, or the invitee chose not to participate. Due to the diverse range of knowledge and experience of the selected participants, interview questions centered on their background, knowledge, and experience in implementing intersectional notions of identity, how their thinking has evolved in the context of rapid technological change, their specific recommendations for smart city initiatives, and any insights regarding transnational or regional change. The study was granted approval by the Australian National University human ethics committee under protocol 2019/732. The experts provided their informed consent to participate in the study and use their full name in this publication.

| Name and Organization | Country/Region of knowledge/experience discussed for this study |
|---|---|
| Diane Bell, Distinguished Honorary Professor, Anthropology, ANU College of Asia and the Pacific | Australia |
| Genevieve Bell, Distinguished Professor, Florence Violet McKenzie Chair, Director of the 3A Institute, Australian National University and Vice President, Senior Fellow, Intel Corporation | Australia |
| Nandini Chami, Deputy Director, IT for Change | India |
| Melissa Gregg, Principal Engineer and Research Director, Client Computing Group, Intel | Asia-Pacific |
| Anita Gurumurthy, Executive Director, IT for Change | India |
| Sue Keay, Research Director for Cyber-Physical Systems, Data61 | Australia |
| Padmini Ray Murray, Founder, Design Beku | India |
| Nimita Pandey, Research and Information System for Developing Countries | India |
| Ruhiya Kristine Seward, Senior Programme Officer, Networked Economies, International Development Research Centre | Asia-Pacific |
| Araba Sey, Principal Researcher, Research ICT Africa | Asia-Pacific |
| Hannah Thinyane, Principal Research Fellow, UN University Institute in Macau | Thailand |
| Amanda H. A. Watson, Research Fellow, Department of Pacific Affairs, ANU College of Asia and the Pacific | Papua New Guinea |
| Joanna Zubrzyki, Associate Professor of Social Work, Australian Catholic University | Australia |

**Table 2: List of interviewees and spread of knowledge/experience across Asia-Pacific**

We carried out structural coding of the interview transcripts to categorize sections of interviews into themes of inquiry (MacQueen, McLellan, Kay, & Milstein, 1998). This is a particularly useful strategy when the research is exploratory in nature (Miles & Huberman, 1994), as in this case. In a second round of analysis, we selected quotations where there was a high level of agreement, difference, or nuanced opinions amongst the experts. We included illustrative examples to give richness to the theme when possible.

### 4.1. The 3A Framework

The themes of inquiry we selected were based on a new framework being developed, tested, and iterated by the Agency, Autonomy, Assurance (3A) Institute, called the 3A Framework. The 3A Framework is structured around six themes, each grappling with a core question to unpack interplay between people, technology, and the environment:

- *Agency:* How much agency do we give technology?
- *Autonomy:* How do we design for an autonomous world?
- *Assurance:* How do we preserve our safety and values?
- *Indicators:* How do we measure performance and success?
- *Interfaces:* How will technologies, systems, and humans work together?
- *Intent:* Why, by whom, and for what purposes has the system been constructed?

The 3A Framework was developed by Genevieve Bell, Director of the 3A Institute, and is based on over 20 years of experience working at Intel Corporation. From 2017–2020, it has been expanded and tested by the staff at the 3A Institute. To date, the Framework has been used and clarified through qualitative case study research, partnership work with industry, and through a series of educational experiments, including micro-credentials and a prototype Masters in Applied Cybernetics – supported by Microsoft, KPMG, and Macquarie Bank – that involves two cohorts of highly-skilled, multi-disciplinary, and diverse professionals. In this paper we evaluate the appropriateness of the

Framework to guide inclusive policy and practice with and for women in the context of AI-enabled smart cities. The following section details our findings.

## 5. Findings

This section outlines the findings of our interviews with experts in relation to the 3A Framework.

### 5.1. Agency: The need to reconstitute AI technology design processes

Across the two cases of AI for social good identified above, there are tasks that can be performed without human oversight. In some instances, women can be personally identified on the street and a prediction made about where they are going or what actions they will take (Huawei Enterprise, 2019). Likewise, in the case of mobility, sensors and cameras, combined with machine learning algorithms, monitor and manage traffic flows. Policymakers will need to work through whether these functionalities are desirable or empowering for women.

A main problem that experts mentioned was that it is often too late to consider what technology should and should not do by the time it is developed and implemented. Experts cautioned for the need to reconstruct the design phase of the AI technologies we considered. Actors need to make explicit the types of problems viewed as being important (or profitable) enough to solve, the underlying assumptions made, and who is included in the process of defining and solving problems. However, there were differing opinions regarding how diverse women should be represented in this process. Genevieve Bell, who grappled with these issues in her role as a Senior Fellow at Intel, explained that:

> "[It's] not just about having more women in the room when the decision is being made. It's about structuring the way the decision is made completely differently because [there's] no point if you're still driving to building a technology in one place and scaling it to the planet. It doesn't matter how many other voices you've got in the room, if they're not the right voices it makes no difference.

*And thinking about who the right voices would be. That doesn't just mean having women in the room. It means having women for whom this might be their community… you have to change the nature of how decisions were made, how conversations were constituted, how you thought about hearing different voices in the room and making room for people, and how you thought about what the logic was under which you are operating."*

We debate the topic of representing women further in Section 5.3. However, what we emphasize here is how the design process of an AI technology might be structured and, ultimately, what technology is meant to do (i.e., the intent behind it. See Section 5.6). Genevieve Bell argues that clarifying decision-making processes and increasing diversity in thoughtful and intentional ways precedes decisions about what AI technology can and cannot do.

There is another thread related to the importance of incorporating intersectional theory (Crenshaw, 1991) into design practice, as articulated by Joanna Zubrzyki, a lecturer in social work from Australian Catholic University:

*"I think it's really important not to also essentialize or stereotype that all women will have the same sets of values just because they're women…. One of the really important contributions I think of postmodern feminism was to say that you just cannot make global assumptions about the lived experience of all women and therefore the values of all women."*

Padmini Ray Murray, founder of Design Beku, a collective working at the intersection of design and technology in India, with substantial experience implementing intersectional notions of identity in smart city design, reflected on how difficult this can be: *"Histories of feminism in this country have been articulated and published by the dominant caste, and so therefore what is seen as "Indian feminism" is kind of seen through the lenses of the savarna woman, who embodies the dominant caste woman".* Approaches to balance dominant voices are also discussed in Section 5.3, yet, here we note how challenging it can be to resolve these sorts of issues.

Design processes need to also incorporate a context-integration phase. Genevieve Bell is wary of the temptation to *"build a global thing and then just have localization strategies"*, and that *"[creating] a series of locally inflected designs that have some common threads"* is more achievable using a bottom-up approach rather than a top-down one. One advantage of such an approach is *"you hear what the genuine set of problems that people feel are, that need to be solved… sometimes that what you think you know about the place isn't what is the problem people want to solve locally" (Genevieve Bell).*

A good example of how design processes can integrate these insights when working on problems of high relevance to women is Hannah Thinyane's Apprise System (Box 1). Thinyane and her team at UN University Macau have been exploring how digital technology can be used to reduce the exploitation of workers in four sectors of employment in Thailand: manufacturing, fishing, forced begging, and sex work. Following a values-sensitive design approach, Thinyane developed Apprise, a multilingual expert system, to support frontline responders (labor inspectors, police officers, community organization representatives) to identify victims of forced labor and human trafficking. Frontline responders access the application on their phone, accessing a question list that has been developed to screen for potential vulnerability:

*"The question is a yes or no question, which makes it easy for us to compute afterwards the vulnerability of the situation there. So, how exploitation looks different in different sectors. So, the kinds of questions I might ask in different industry sectors, say in fishing, 'I might not let you off your boat', when you're in port, that's a way of confining you. And in sex work it might be that I won't allow you to choose your own customers."*

According to Thinyane and Bhat (2019), there is a gap in our understanding of how many workers are often placed in challenging situations that are not clearly forced labor, that have begun on consensual and mutually beneficial terms but devolve into abusive work relationships. It takes strong community

relationships and cultural sensitivity to be able to tease out whether a worker is vulnerable or not. For instance, in the sex industry, Thinyane explains that the question list was developed in consultation with sex workers, community-based organizations (CBOs), and human rights lawyers. It also incorporates empirical evidence drawn from research with over 3,000 sex workers to identify the top four practices of exploitation within sex work in Thailand (Empower Foundation, 2012).

Another approach emphasizes women's empowerment, rather than focusing the design process on addressing specific problems. Empowerment broadly refers to capabilities to control one's life choices or the decisions that affect one's life, with the literature defining numerous dimensions and structural aspects to consider (Friedmann,

1992; Oakley, 2001; Perkins & Zimmerman, 1995). Anita Gurumurthy and Nandini Chami are from IT for Change, a leading civil society organization headquartered in Bangalore, India. IT for Change is engaged in research, policy advocacy, and field practice at the intersections of digital and data technologies, with social justice and equality at the international, national, and local levels. Their approach to women's empowerment rejects one-size-fits-all solutions, enabling women and girls to define what empowerment means for themselves:

> *"The team that works in schools has sought to build a curriculum that uses the Internet and digital media to create spaces for self-reflection and collective reflection among adolescent girls, where they can chart out their own definitions and descriptions of what it means to become empowered through technology."*
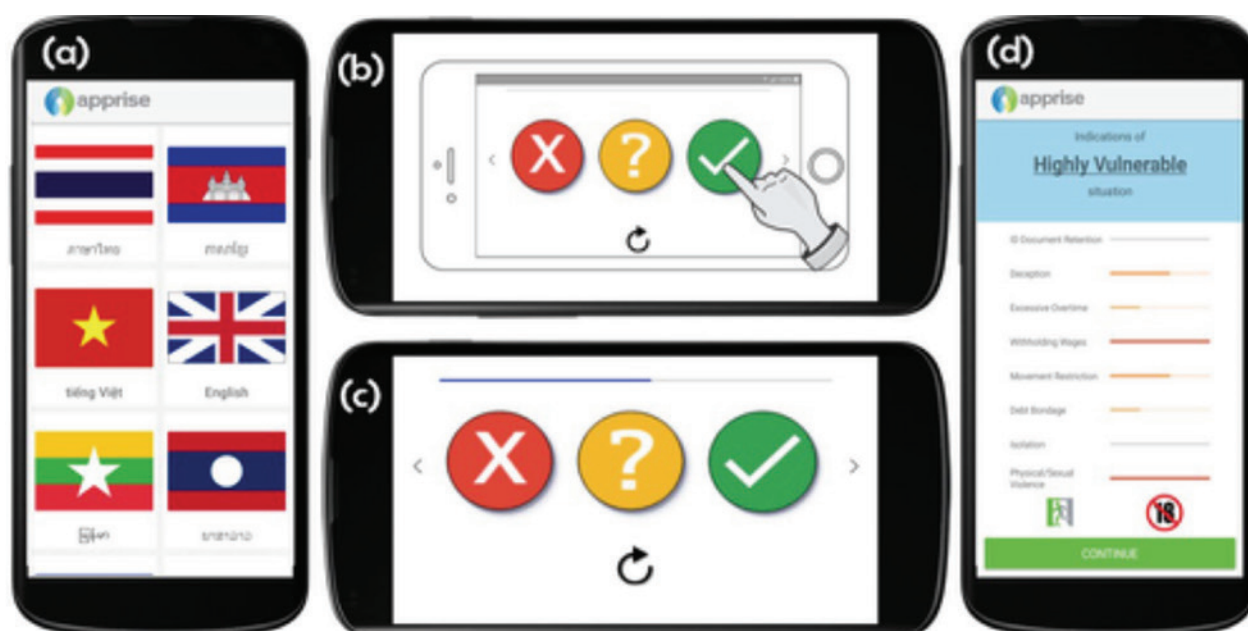


As pictured in (a), a frontline responder will give a smartphone to a worker to select a language. A series of questions are spoken to the worker in their own language whilst they are wearing headphones, so that they can respond without scrutiny of the responder or a translator. Thinyane notes in earlier research that translators were often not trusted or corrupt. Workers likewise felt embarrassed to answer the questions honestly out loud to the responder. Figures (b) and (c) show the interface that workers see when answering the questions. Once the worker answers all of the questions, they hand the phone back to the responder and it displays a categorization of the worker for the responder to review (Figure (d)). They may then offer additional options or avenues of support. In the future, Thinyane's team hopes to use Apprise to identify patterns of exploitation, which machine learning algorithms may facilitate.

The premise that individual reflections should factor into the design decisions of AI technology is complex, indicating that a new area of research is warranted. However, protecting and supporting such spaces for reflection is also important when it comes to enabling women's participation in AI-enabled smart cities:

> *"I think that what AI is going to do for women's empowerment, and what that would mean for the ideal of gender equal or feminist AI futures, is not only about women's safety in smart cities. It's really about the idea of that city in terms of many different citizenship planes… If you look at AI as being integrated into a larger economic ecosystem, or AI also re-architecting these larger economic and social systems, we see that AI becomes part of that important ingredient which is in a dialectic with society, policy, politics, and economics"* (Anita Gurumurthy).

Gurumurthy stressed that attending to how AI technology contributes to women's participation in different "citizenship planes", is crucial for women's empowerment. IT for Change has therefore confronted these issues by researching critical pedagogies, capacities, and impacts, which is then incorporated into their advocacy and policy work, and filters back into their practice engaging directly with communities.

Overall, we find that there are a number of conditions that need to be addressed in the design phase of AI technology before a discussion can take place about what AI can and cannot do. This reinforces the need to simultaneously investigate issues across the 3A Framework (see Sections 5.2 and 5.3 in particular). However, what we conclude is that inclusive AI requires greater attention and transparency regarding decision-making processes surrounding its development, particularly when identifying problems and intended outcomes. There were differing opinions regarding which actors should be involved in decision-making surrounding design decisions. Some experts believe strongly in the need to incorporate participatory democratic technology design processes, underpinned by women's empowerment objectives. At the very least, there is a need for designers of AI technology to

generate stronger links directly with diverse members of a community, as all the experts interviewed supported intersectional notions of identity.

## 5.2. Autonomy: Building for the diverse realities of women

Designing for an autonomous world will involve being sensible to the practical realities women face. This involves many facets of a woman's life and various aspects of her identity, not just those directly implicated in an AI-enabled system. As outlined in Section 5.1, one of the points of agreement amongst the experts interviewed regarded the importance of context-driven feminist praxis: "You're talking about a region that is essentially both on the infrastructure side, socio-economic development side, as well as the digital innovation side very, I would say, highly fragmented. You're not really seeing one picture. And therefore, there can be no one-size-fits-all" (Anita Gurumurthy). This theme focuses on understanding what processes and relationships AI is automating, and what they reveal in terms of the power and position of women in smart cities at the time. There are three key insights from the experts concerning the impacts of underlying infrastructure, differences in access and abilities across marginalized populations of women, and the need to educate women about the changes AI introduce as part of the process to include women in smart city development.

Our literature review shows that it is common for Asia-Pacific countries to pilot smart city initiatives in places where it may be easier to implement AI-enabled systems in terms of the required underlying infrastructure. Experts discussed how exclusions can take place on three levels: the country level, city level, and within cities. At the country level, countries lacking in power and Internet infrastructures exclude many AI for social good applications. Amanda Watson, a Research Fellow that has been researching mobile phone use in Papua New Guinea for 12 years, provided a useful criterion for systems-development there:

*"Every single time someone mentions a possible project idea to me, which has happened many times over the years, I frequently say, first of all, can it work offline? So, if there is an Internet or cloud element, can it still function if you have no Internet? For instance, can the information be stored locally… what's the battery life and power if there's some sort of device because electricity does go down."*

Many citizens in Papua New Guinea do not have electricity, and the electricity grids that do exist may depend on solar energy. Indeed, many remote Australian towns and cities face similar constraints.

Exclusions also happen at the city level, as Nimita Pandey, a Research Associate working for New-Delhi-based Research and Information System (RIS) for Developing Countries, with expertise in science, technology, and innovation policy perspectives, described regarding the choices India has made:

*"There is a huge list of criteria and processes that they opt in picking up cities, in order to make them smart. And the idea of making them "smart" is to make them "sustainable" in terms of energy, in terms of infrastructure, in terms of quality of living. But while doing this, the idea of "sustainability" is lost; it actually causes "exclusion". And this exclusion is not merely from the gender perspective, but in terms of the socioeconomic demographic angle as well. Most of these smart cities are not accessible to everyone who is part of the city."*

Pandey argued that women are also excluded in heterogeneous ways within cities. In infrastructure-poor locations, smart cities may need to either decentralize management of autonomous systems or find specific ways to include marginalized women. Anita Gurumurthy, in reflecting on the Indian context, felt the latter was critically important:

*"And so smart city data for energy management or water management or housing, each of these is not going to be managed in silos. The city will manage all of this data in an integrated way and therefore it is basically a question of where are women in a participatory democracy? Is the*

*data management system reflecting their concerns? What is it that women have to say about water consumption in the city? Which women's voices are being captured by the system? Is it covering the voices of the women who are waking up early in the morning to fill their pots in these slums and then rushing as domestic help to work in somebody's house? And struggling to send their daughters to school whose safety they can't ensure? And also finding city transport, creaking under the pressure of efficiency."*

Nandini Chami likewise felt that new models of ownership are required:

*"We need to think deeply about the design of smart city projects. In the data systems being set up in these projects through public-private partnerships, who should be the trustees for the management of common data resources? Can we assume private companies will automatically uphold public accountability or do we need completely new arrangements for the stewardship of citizen data? We need a radical overhaul of data governance frameworks."*

The ownership of data resources is a particularly sticky topic where AFRT and mobility pattern recognition are concerned, which is further discussed in Section 5.3. However, Chami gave the example of South Korea, who opted to create its own mapping platform to map various resources (not just locations) (see Korea Legislation Research Institute, 2019). This strategy may support Asia-Pacific countries to adopt heterogeneous models of integration for autonomous systems, which could address the needs of diverse women. In South Korea's mapping platform case, contributing actors need to be able to frame their service in terms of the platform aims and how the benefits would be shared publicly. This would encourage companies to make explicit how their service responds to particular populations of women in a specific context.

A second aspect that needs consideration relates to how diverse women have differing abilities to both understand and interact with automated processes and technologies. There is limited empirical evidence

regarding how women across Asia-Pacific might have different capacities to engage with CCTV or smart public transportation systems. We do know from experience that marginalized women have drastically different needs stemming from divergent cultures and capabilities of interacting with technology, such as smart phones. IT for Change has been supporting women across rural and urban settings in India, investigating how technology can be used to empower girls, all the way up to elderly women. They have learned to adapt their engagement strategies to various levels of digital literacy and technological usage patterns. Upon speaking of a project based in Mysore with older women, Chami mentioned post-literacy approaches for empowerment education: *"for example, you cannot use a lot of text-based aids or learning materials. One would have to rely a lot more on highly audio-visual tools: videos, digital stories, [and] voice messages on mobile"*. This implies a necessity to account for skill and cultural diversity when embedding automated technologies and processes into an environment.

Padmini Murray, having conducted one of the only studies on the experiences of girls in the smart city, also found that girls in Delhi were reporting new risks that required mitigation: *"I think what was most visible was that patriarchy enacts itself through digital vectors as well as through the material. So, you would have things like girls complaining about being sent pornography, harassment on platforms themselves"*. Melissa Gregg, along with Genevieve and Diane Bell, likewise expressed the importance of ethnographic fieldwork as a means to understand the particular challenges experienced by women as new autonomous technologies are introduced. However, Gregg cautioned that at times it may not be obvious what processes AI is automating: *"What I wonder though is… how much do people even know about what's being collected about them right now. So, [for] me, my first question is how are people even made aware of how they are tracked?"* As discussed further in Section 5.5, Ruhiya Seward argued that new education programs are needed.

In sum, whilst AI technology and the integrated systems needed to implement these technologies into Asia-Pacific cities is important, our research emphasizes that inclusive practice comprises three aspects. Firstly, when taking autonomous processes and systems to scale, policymakers need to make clear links between plans to reduce infrastructure inequalities and plans to develop smart city initiatives. Secondly, regardless of successful pilot tests, gender and cultural diversity are clearly factors that will impact on the roll out of autonomous systems. Greater attention and planning must be paid to accompany implementation through research and refinement to customize and problem-solve across contexts. Thirdly, citizen education programs are urgently needed to raise awareness of the myriad impacts and implications that automated processes have.

## 5.3. Assurance: Ensuring diverse women's needs and values are heard

Assurance refers to the practices, processes, institutions, and rules that ensure the safety and respect of societal values, especially from the perspectives of diverse women in this case. It is therefore not only structured by one relationship but by a system of relationships between all actors involved – including AI technologies and systems.

If assurance is conceptualized as a system of relationships, the experts interviewed have worked tirelessly and consistently to ensure that women are key actors, whose voices have a right to be heard in such a system. The main difficulty in the context of AI-enabled smart cities is the lack of clear roles and opportunities to participate in decision-making surrounding how these initiatives impact on women's lives. There are lessons to be learned from the struggles that the experts interviewed have confronted in their own lives and careers. For instance, Diane Bell, acclaimed Australian feminist anthropologist, recounted the struggles she endured to pursue her education, and to gain access to scholarships and grants as a single parent:

*"I was the first woman to do anthropological fieldwork in Australia with two children as a single parent. There had been women in the field, but as a wife looking after his children, or it had been a woman just for a very short period or somebody had taken the kids. All the major women who'd worked in the field were single and had no children."*

Moreover, the experiences of the experts also highlight what it means for women to claim greater accountability for the conditions and quality of life imbued by AI-enabled smart cities. As Diane Bell expressed, concerning her experience working with Aboriginal women in Australia:

*"How do we get all those voices to the table? How do we hear from those people? How do we make the conditions so that all of those are there? But why should it be "we" making the conditions? How do we have it so that those people are saying, "This is my issue too". … How do we get that consciousness of who's at the table? To understand how these broader issues are interrelated? An "Aboriginal issue" is not just about where do I live because I'm Aboriginal and how is my language and my culture respected, but why am I not at the table on issues of national security for instance? Where should my understanding and my history be understood? [And] it should be right across the board."*

Sue Keay pointed to ethics panels, especially in a medical context, as a good example of consulting with people who are representative of a diverse community. Joanna Zubrzycki talked about her work with indigenous people and noted that you cannot always get everyone to the table at the same time, so *"[you've] got to reach out and ensure that you are listening… and find those diverse perspectives… you've got to make the effort to go to people to consult".*

Yet, in the current phase of technological development and implementation, we have seen limited evidence of consultation or participation in decision-making, reducing the scope of effective local governance of which Diane Bell speaks. Therefore, the experts speculated about mechanisms that may return

attention to questions and issues of participation in smart city governance. One area that emerged regarded data ownership and governance. Araba Sey explained:

*"What should happen, or what might be more practical, is for government and civil society organizations to find ways to partner somehow with the commercial or corporate entities to ethically get access to the data that they automatically generate, and try and use it in ways that go beyond just making profit. That may be an arrangement that could possibly at least share the responsibility, and make sure that it's not just the corporate bodies that have access to the data and use it only for economic gain."*

In contrast, Anita Gurumurthy reflected on their experience developing a community-based water management app in Bangalore, India; and the steps taken to enable collective ownership of data and the skills needed for citizens, women, and men alike to use the system to claim greater accountability from local officials:

*"This is the idea of [a] smart city that we think should really be replicated, not necessarily to scale in a homogenized fashion, but in context-appropriate ways based on the particular needs of communities. There should actually be a way by which communities can manage their data and engage with local authorities for claims-making, with the complete knowledge of how data interfaces work."*

However, as Padmini Murray pointed out, referencing Baud et al. (2014), the ways in which similarly participatory democratic processes have been implemented in smart city initiatives has tended to over-index the perspectives of the middle-class, leading to significant bias in interpretation and inclusion. To work towards *"resolving intersectionality with consensus"*, Murray, along with Mozilla Fellow Divij Joshi, are developing an automated decision-making system precisely for this purpose. They are constructing an interactive platform that *"demystifies how automated*

*decision-making is done in the smart city. What technologies are used, what is the data that those technologies [are using], what are the assumptions, rather, that are being built into those technologies to take the decisions that they do".*

Another essential intervention strategy is to significantly increase evaluations, including social audits of AI-enabled smart city initiatives. As Anita Gurumurthy argued:

> *"About four years ago, after the very unfortunate event of a young woman student in Delhi being raped, a fund was set up by the government, and then UN women and many other actors then got on board to initiate action on women and safety. Many apps were introduced as part of such action and I'm not really sure whether the assessments and evaluations of these really do exist. I haven't seen many. We work on the whole idea of feminism in technology and I do think that we should really be having many more evaluations."*

Whilst we elaborate on potential purposes of evaluations in the next section, generally speaking, the assurance theme highlighted that voice, representation, participation in decision-making, and community ownership are of great consequence to including women in AI-enabled smart cities. There is reason to explore innovative ways to address these processes and topics, as Murray is doing. Indeed, this thematic area seems critically important to empirically research further.

## 5.4. Indicators: Addressing root causes rather than symptoms of gender inequality, and the concept of equity

When AI technologies are embedded within urban infrastructures, they may be designed and evaluated with a certain purpose in mind. Measuring the performance of a remote sensing system for traffic flow management might focus on indicators related to time or congestion. Likewise, facial recognition systems might also monitor error rates and positive identification rates. In either of these

cases, performance measures emphasize envisioned purposes of technology and their overarching efficiencies. Nevertheless, these technologies affect critical infrastructure and the social fabric within which urban living takes shape. Moreover, the inclusion of women in this context implies gender relations will be rebalanced in the process. Yet, the needs of diverse women and men are complex and are particularly challenging to measure.

Gender inequality observed in both access to technology and its related industries are the main challenges to which the experts are no strangers. Women tend to have less access to technology across four basic access indicators: computer use, mobile phone ownership, mobile phone use, and access to the Internet (Sey & Hafkin, 2019). Women also constitute less than 35% of information and communications technology (ICT) and related professions, with substantially fewer in leadership positions (Sey & Hafkin, 2019). It is this persistent awareness of the severe gendered imbalances in access and usage patterns, affordability, workplaces, and industry representation that propel experts to engage in generating knowledge and praxis to bridge divides. Araba Sey is a scholar who has worked for the last three years on the UN's Equals in Tech initiative. Prior to that, she investigated inequality between nations in terms of ICT infrastructure and uses, as well as between socioeconomic groups within countries for more than a decade. As someone who understands these imbalances all too well, Sey expressed frustration regarding how our knowledge of the issues points to little progress towards resolving inequalities:

> *"I feel like some of the things we're measuring need to start at a much, much earlier age, and may not all be as quantitative as the current trends in the collection. I feel that a lot of what happened could be addressed at early stages, so at the primary elementary school level and then in the home, so that things like [a] parent's attitude towards gender… or towards [their children's] career [choices]."*

Sey's advice is to concentrate efforts on addressing the root causes of gender inequality rather than treating symptoms down the line. However, although it might seem out of scope to address gender inequality issues within smart city initiatives per se, it could be an important mitigation strategy.

In contrast, Diane Bell spent decades researching and advocating for Aboriginal Australian women. Her experience highlights how the international framing of gender equality within the Sustainable Development Goals (SDGs) may not be an appropriate standard to set. On speaking of her fieldwork from the 1970s in central Australia:

*"They had very independent lives. They hunted and gathered for their own food. Some of that food would go to their menfolk, but they were self-sufficient in themselves. They had their own camps that were organized according to their relationships to country and they had their own ceremonies which were organized by themselves... The notion that there was a feminist perspective on practices that might be underwritten by shared values and principles but were pursued in separate spaces was very clear to me. And that was a difficult thing to explain within*

*the white women's movement at the time, which wanted equality and integration. And I was saying there are other models. There are models with independent bases of power and standing."*

Alternative models (to equality) based on independence and freedom to define one's measures of success is similar to IT for Change's approach to women's empowerment discussed in Section 5.1. Both require sufficient trust and time to establish as a means to protect "independent bases of power and standing".

Trusting relationships are indeed critical to developing measures of success shared across organizations. Ruhiya Seward, based in the Amman, Middle-East office of the International Development Research Centre, and working on the technology and innovation area in the Networked Economies group, has been working to improve gender-related outcomes across her team. She has also been overseeing feminist projects including the Gender and Technology Network, led by the Association for Progressive Communication (APC). She reflected on the specific challenges of working collaboratively across institutions:
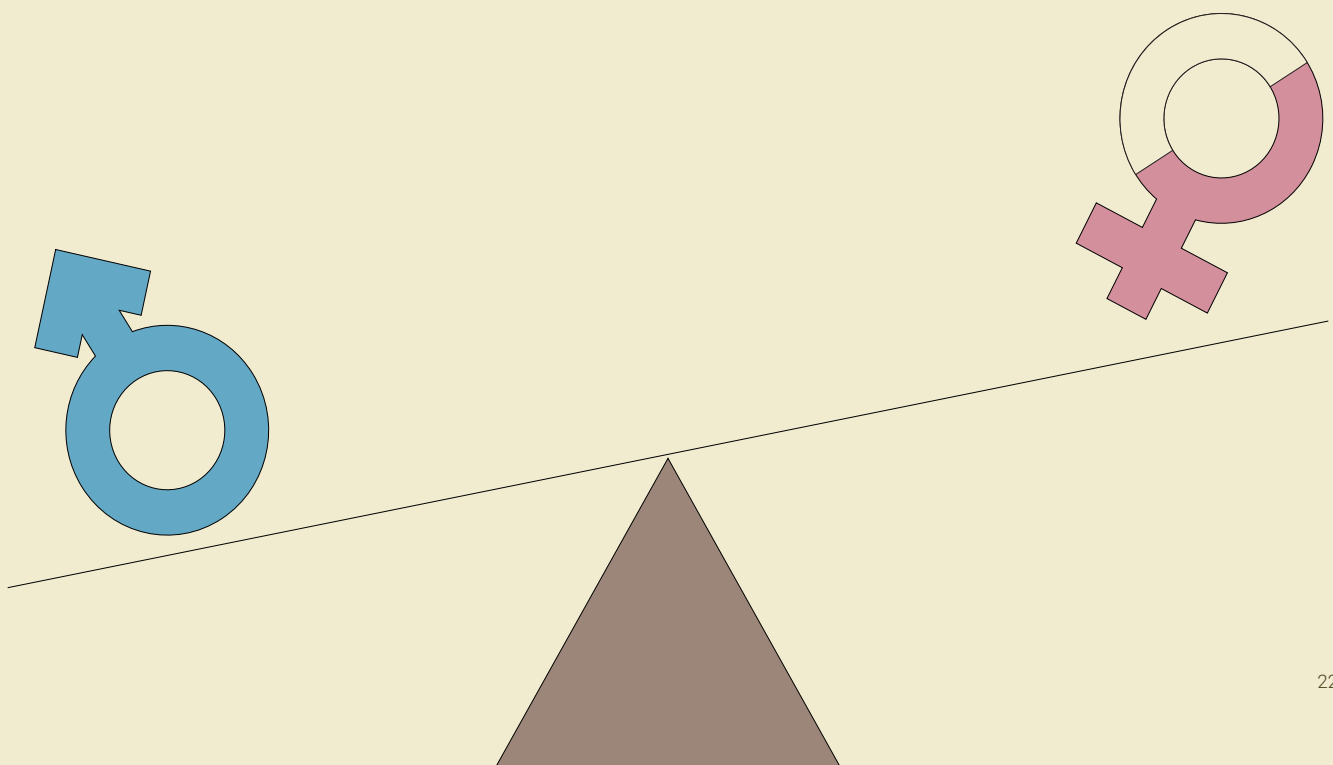
"Feminism is, in a way, depending on how broad your umbrella is, what we might call kind of participatory democracy or even democratic socialism. It takes time to activate. And yet there are the realities of getting work done and being responsive and doing stuff and forging forward and having a strategy – these challenges don't always lend themselves to an amoebic participatory/collaborative management…. This can be a challenge when it comes to policy ecosystems versus feminist ecosystems… You actually need policy outcomes in order to show that it's valid and worthwhile and that you're spending public money in good ways."

There may be some indicators that can be negotiated, whilst others cannot. This may be why it is also beneficial to establish shared principles of success. Nimita Pandey's organization, RIS, developed a framework to contextualize responsible research and innovation (RRI) in India. She mentioned that the framework provides a principled basis to examine the social dimension, spanning multiple projects and contexts:

"From a developing country perspective, we proposed the [Access, Equity, Inclusion] framework… because while reflecting at gender under the project(s), it has emerged as a very critical issue; even there have been mandates across different departments, particularly the Department of Science and Technology. Studies would definitely add to our methodology, in order to develop an exhaustive list of indicators to assess and evaluate programs and initiatives, in order to find the enablers or barriers, which are critical for gender inclusion."

Initiatives such as these could potentially be integrated into smart city projects as a means to monitor and evaluate gender issues across projects.

Lastly, many of the experts agreed that including women in AI-enabled smart cities depends on the participation of women in the relevant skilled professions, policy spheres, public services, and leadership roles. Sue Keay is the only female research director (of three) at Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO). She leads four group leaders, none of which

are women. CSIRO joined the Science in Australia Gender Equity (SAGE) program, which is a partnership between the Australian Academy of Science and the Australian Academy of Technology and Engineering. Its vision is to "improve gender equity in STEMM in the Australian higher education and research sector by building a sustainable and adaptable Athena SWAN model for Australia" (SAGE, 2018, n.d.). Such a model provides a charter of principles to ensure that their policies, practices, action plans, and culture reduce gender inequality. Sharing data on these matters enables some accountability for this issue from the organization. However, Keay felt that a great deal more needs to happen:

"With these initiatives that I personally was following or kind of I was asked to do, I guess unfortunately, they're all things that I've decided to do. I would prefer if that was just a priority for the area that I work in, but at the moment it's not… I'm increasingly feeling that it actually has to be something that is mandated, that it's compulsory that there is no ifs, buts, or maybes, people just have to do it. And it doesn't actually matter the reason, people [just] know that they have to think about safety in the workplace, they should also have to be thinking about inclusion in the workplace… I certainly believe we must be publishing metrics."

In sum, indicators designed to establish and track progress towards various levels of reducing gender inequality within AI-enabled systems are needed. The experts flagged three scales of complexity to consider: firstly, indicators relating to global gender equality targets (or alternatively, independently defined targets); secondly, indicators relating to specific projects or programs; and, thirdly, evaluations must seek to uncover how AI-enabled smart cities address the root causes, not only the symptoms of gender inequity.

## 5.5. Interfaces: Defining boundaries and considering accessibility

As outlined in Sections 5.1 and 5.2, the AI-enabled smart city technologies we consider are rarely designed in a manner that aligns with the feminist praxis discussed in the interviews. The experts

identified how interfaces within AI-enabled smart cities are a crucial element to consider where diverse women are concerned. The most-marginalized women within Asia-Pacific cities are potentially concealed and further disadvantaged when they lack the accessibility and knowledge to interact with the interfaces of a system. Women's public spaces are also increasingly occupied by the sensors and cameras needed to operate smart traffic systems and CCTV systems. This occupation has implications on the definition and communication of boundaries to acknowledge where interfaces begin and end.

Regarding accessibility, Ruhiya Seward remarked that severe access inequalities necessarily impact on how people may experience interfaces with a system: *"So many people in the world don't connect [to the Internet] at all, which means that they don't show up in the data. If you don't show up in the data, you don't matter to AI"*. On the other hand, by building AI technologies into city infrastructures, women may have less of a chance to decide whether or how to connect with a system. Hannah Thinyane referred specifically to this point:

*"What if they don't actually have an ID document? What if they don't want to be known? There are all of these things you have to consider when designing a system that will be citywide. I guess how does it also work with disabilities? How does it include disabled people? And then with migrants, and Thailand has such a huge population of migrant workers, how have people (documented or undocumented migrant workers) [been] included in a design of a system like that? Especially if it's got anything to do with identity."*

If actors managing AI-enabled processes do not incorporate inclusive practice, there may be no way to tell if interfaces with a system actually function, or are desirable for diverse women.

The occupation of public spaces by new interfaces with AI-enabled systems is also a concern. Melissa Gregg reflected on some of the challenges emerging from her involvement in the research and development of smart home devices, primarily in the US:

*"One of the things that really struck me, [for example], is how services like Amazon Alexa, the Echo, and other technologies were being brought into the home with a very gendered voice. As a sort of idea and subservience that is very familiar for women in domestic environments. Having that background let me think about what is being normalized by the design of these devices. But then as the ecosystem developed towards Amazon's ties to the Ring doorbell, for example, it made me stop to think about the role of the household within a neighborhood… It really worried me when I started to realize that Ring had arrangements with local authorities in certain neighborhoods that there was subject to screening from some of those law enforcement officials. The idea of the state in the US again is a little dis-aggregated from your local street. So, [for] me, that's a clear example of how if there is a thread of how the woman at home is under threat and technologies are designed to enable a certain kind of efficiency of monitoring, whether they're in that home or outside of its perimeter. I don't know that is the thing that concerns me a lot, which is what has been traditionally gender roles of care and nurturing and support and community relations becoming instrumented in these data gathering devices."*

Gregg directs us to some of the more entrenched impacts of integrated services combined with AI-enabled devices, and how women's voices may suggest care and nurturing, yet the involvement of law enforcement may be otherwise experienced. Whilst she noted differences in relations between women and the state based on her experience working across the US and parts of Asia, such as China, Japan, and Korea, what is important here is the capacity for women to be embedded in very seductive, or what Gregg calls "normalizing", activities enabled by AI in the name of well-being, without understanding when these devices are interfacing with new sets of actors, such as local police. Moreover, Padmini Murray's research, with Prof. Ayona Datta, uncovered how in India, when young women chronicled their engagement with the smart city in Delhi through daily WhatsApp diaries, they often found it difficult to draw boundaries between their experience with the city and "the smart city":

*"So, I think we found that they would often tell us about ways in which the infrastructure of the city would let them down. During the monsoon, Delhi would flood very easily and how that would cause a lot of difficulty, even would cause deaths because of electrocution and things like that. So, the picture that we got from their journals was just basically that they were always at war with the city. But what wasn't immediately available to us was how does the smart city impact... It's not really possible for them to parse what the city is doing to them through the lens of what the smart city is doing. And it also depends on what we mean by the ravages of the smart city."*

In this case, interfaces are often difficult to identify or disentangle from the broader city living experience. The majority of the experts interviewed argued for greater transparency and education opportunities to help women understand and claim their rights in this context, as Ruhiya Seward stated:

*"Most people just don't really understand data ecosystems. They just don't have a fundamental understanding of their own human rights, of what data can do… Inclusion is having the skills to know what your rights are, and activating those rights, and working with them."*

The interfaces theme draws out concerns about whether AI-enabled systems have designed interaction experiences for diverse women, especially the most marginalized who often lack accessibility to technology that is used to gather data for AI. More importantly, there is a need to make the interfaces of a system visible and to debate the terms of informed consent in this context.

## 5.6. Intent: Examining power relations and potential misuses

A central concern raised by the experts reflects the dual nature of AI technology used within smart city initiatives. Even if facial recognition can be used for stated purposes related to safety and security, it enables other outcomes that may be experienced as harmful, such as increased surveillance and control, lack of freedom of expression, and unknown data privacy management practices.

Diane Bell spoke of the need to question what problems AI is meant to solve, and having the capacity to debate whether or not it serves the collective interests of citizens, including those of diverse women:

*"AI has enormous capacity to improve our lives, but is it being developed within a framework where the narrative is one of rights and responsibilities, or is it developed because we can do it, therefore we'll do it? Not, why should we do it? Well, we can do it, but should we do it? There's many things we can do but should we?"*

Genevieve Bell spoke of the reality underlying the development of many smart city initiatives:

*"So, if you imagine that most technical systems are not built because someone has a generous whim, they are mostly built because they are either designed to perpetuate power, or general capital, or both… So, it's not surprising in that sense that most technologies sit within systems of disenfranchisement because that will be the flip side of power and money."*

These quotes challenge policymakers and practitioners to expose power-relations within a system, and to ensure that the intent of AI is balanced by a framework of rights and responsibilities.

Furthermore, almost everyone pointed to the challenges of acknowledging intersectional differences in power and access in context, where the intents of the more powerful or directly implicated are at play. Genevieve Bell gave the following example to highlight this point:

*"The classic example for me about the place that went horribly wrong... might be Chicago, certainly Illinois… [where] they had a smart traffic lights system… [that] was being run not by the police but by an outside third party. And in order to hit their revenue targets every quarter, they used to vary the traffic signal rate. So, the amount of time the light was yellow used to diminish towards the end of the quarter so they could catch more people running red lights."*

She argued that we need to ask questions that are not necessarily about gender but about the problems that the system is intended to resolve: *"How do you start to imagine what is safe, right? Because what a government decides is safe may not be what its citizens decide is safe."* Joanna Zubrzycki agreed, noting that it is often in working through the intent of a system that policymakers may begin to deal with the complexity of AI-enabled systems:

*"I mean if you look now at the sort of issues or just last week with the tragedies around domestic violence which people are starting again to grapple with. I think when people start naming those different problems, those intersections become very clear. And I think that's when policymakers start to realize that they've actually got to deal with multiple dimensions of the problem and women's experiences."*

Alternatively, in some countries, the powerful classes of actors may have little power to define their own intents and purposes. As Amanda Watson explained:

*"Many of the Pacific Island nations do have donor funding or if you count the donor dollars themselves going in. It's a huge percentage of the overall budget or the overall money that's spent in these countries. So, I guess that's why or one of the reasons why so many of these things would end up being donor projects, because the governments themselves don't necessarily have money to even run their health and education systems."*

Power-relations are essential to unpack the intent of AI-enabled systems, as well as to situate actors within them and their capacity to address core issues. Working through issues surrounding intent may also gather insights into potential misuses of AI. Ruhiya Seward's thoughts encapsulate comments from a number of experts:

*"I mean essentially, it's kind of a big brother issue, and I don't see any other way of framing it… I think actually this really speaks to the tension of technology in general, broadly considered, in that there are all these potential advantages (and disadvantages) that come with security. [Say] a woman is harassed or attacked. If you have big brother surveillance, it can identify the attacker, track*

*them down, and ensure he or she is brought to justice. That improves the lives of people vulnerable to harassment. On the other side of that security, if you have a state that doesn't believe in free expression, this same technology can be used to track down people who are dissenting, who are protesting, who might not want to be identified or singled out… We know that [democratic systems are] being threatened all over the world… So how do we grapple with this big brother that's here, that's arrived – where we want safer cities, but we don't want our freedoms curbed. Basically, it seems like it's a trade off right now."*

Seward's framing of the multifarious intentions that are purposeful and emergent within AI-enabled systems suggests that potential harms, specifically to women, are vital to evaluate. As Araba Sey related:

*"How do we ensure that those that do have access might not abuse them… This is more about those that have access to the system and ensuring that they are ethical, or… that there are measures in place to ensure that the potential for [misuse] is limited. Because women tend to be the predominant victims of abuse, I think, it becomes definitely a gender-related issue. Women and people of other non-masculine genders tend to be the ones that are victimized more often, so I think there's a definite gender component."*

There are likewise many components and levels of an AI-enabled system that must be considered. Hannah Thinyane spoke about how her design decisions ripple throughout a system, of which they can be taken advantage. Her thought process was:

*"If we captured this extra information, how could that be abused? So, for example, we were asked from very early on if we could capture a camera photo, because say the NGOs would say, if you think of workers from Myanmar, they all have the same name. And if you have five people who were on the boat and they all have the same name, how would you know which one you talked to? … Any system that has corruption, if someone can make a few extra bucks and they don't feel like [they're] paid enough, well they might give that information to someone else."*

Altogether, the intent theme captured the experts' attention to power-relations in context, as well as how these are expressed. When considering the power and position of women, strategies to hold powerful actors to account and to protect against the misuse of AI-enabled systems are needed.

# 6. Securing voice and recasting participation: Examining roles and responsibilities for the inclusion of women in AI-enabled systems

The following two sections discuss our findings and generate key policy recommendations for the roles and responsibilities required to include women in AI-enabled smart cities. We also review our exercise of elaborating on the 3A Framework to address inclusion concerns and reflect on its application as a tool for future policymaking in this area.

## 6.1. Increasing voice and participation of women in smart city initiatives

The past decade has seen growing support for the notion of "inclusion" in the rhetoric of smart city initiatives, yet key decisions that affect women's lives continue to be made without adequate consideration, consultation, or differentiation, especially when it comes to diverse women across various sections of society in the Asia-Pacific region. Why has the rise in the rhetoric of inclusion not coincided with greater scope and attention to the voices of diverse women, especially the most marginalized? How do AI for social good applications change the methods and practice of participation? The rise of AI has occurred simultaneously with some advances in methods and approaches designed for greater citizen engagement in smart city initiatives, such as deliberative decision-making, citizen juries, and public consultations. There is, however, limited evidence that these approaches have been rolled out extensively, internalized, or that they have influenced wider policy or programmatic budgeting and decision-making within AI-enabled smart cities.

As Joanna Zubrzycki poignantly stated, the inclusion agenda risks essentializing women, and can be used to disempower women as much as the reverse. Padmini Murray reminded us that some forms of participation can actually widen the gap between "inclusion" and "exclusion" when certain classes of women are favored over others during consultation processes. All of the experts were likewise in agreement that most women lack knowledge to engage in data ecosystems that underpin AI applications. In Section 5.1, Hannah Thinyane highlighted that vulnerable women are also hesitant to share their perspectives when trusting relationships are lacking. It seems clear that a main purpose of inclusive practice is to support the most marginalized women in smart city design and implementation.

Intersectional feminist theory (Bhavnani, Foran, Kurian, & Munshi, 2016; Crenshaw, 1991) has provided a language to understand the social, cultural, and economic factors that influence the power and position of the most-marginalized women in relation to others, including men. Although all of the experts endorsed this framework for understanding a woman's power and position, it remains challenging to adopt in practice. Examples discussed by the experts incorporating ethnographic accounts (Padmini Murray, Melissa Gregg), participatory models (Anita Gurumurthy, Nandini Chami), and values-sensitive design (Hannah Thinyane, Genevieve Bell), strengthen understandings of women's realities as multi-dimensional, intersectional, and dynamic. These methods may facilitate the inclusion of women's voices in large smart city projects. However, disjoints between rich accounts of women's experiences and the design of AI technologies and smart city infrastructures are still common.

Why is it that intersectional feminism has not entered the mainstream in terms of framing and delivering public services such as AI-enabled public transportation and CCTV systems? Typically, the needs and aspirations of the most marginalized have been served by specialist bodies and organizations, such as social workers, community-based organization representatives, and care workers. A

promising solution might be to educate these front-line workers on the opportunities and risks afforded by AI-enabled systems, and to support their roles as advocates to move this agenda forward. As co-author and a trained social worker, Brenda Martin (2019, p.6) wrote:

> *"In Australia, social workers are often in a unique position to witness the impacts of new socio-technical systems on the lives of our most vulnerable individuals and communities, to analyze structural inequities, to educate, to elevate the voices and experiences of those excluded from public debate, to influence public policy, and to advocate for change. As social workers, we need to develop the language and understanding to be meaningful and powerful contributors to the debate on the current and future roles of AI and cyber-physical systems."*

Such workers and organizations can provide critical questioning and feedback into a system to highlight specific and systemic biases and risks of AI technologies. That said, this policy alone may place greater stress and pressure on an already over-worked professional base, which may spread their responsibilities for women too thin. In the next section, we consider how else to build responsibilities for the protection and empowerment of women into AI-enabled systems, and what these roles and responsibilities might look like.

Moreover, it is not likely that increasing participation of women in smart city initiatives through deliberative decision-making, citizen juries, or otherwise will be enough in the context of AI-enabled smart city initiatives. Particularly in the cases of using AI to increase safety and mobility of women in smart cities, there will be difficulties in establishing the trust and close relationships necessary for an open discussion to share their views and preferences with authorities. Seemingly endless histories of violence against women and social control of women's behavior exists in most contexts across Asia-Pacific. It seems dubious to suggest that women's participation in decision-making processes would be valued and embraced. It also takes time to experience and reflect

on how AI developments will interact with power-relations, attitudes, and behaviors in context. Whilst participatory democratic processes should certainly be prioritized, the costs and technical expertise required to implement many AI-enabled smart city systems puts pressure on authorities to ensure strategic returns on investments. There is still a need to develop checks and balances, along with rewards and incentives within a wider network of smart city actors.

## 6.2. Roles and responsibilities in an interlaced network of actors: The value of applying the 3A framework

This research elaborates on the 3A Framework as a tool to outline the contours of inclusive practice within AI-enabled smart city systems, whilst taking into account the culture and values of diverse women. Our review of the literature and analysis of the interviews with experts, points to the key issues that the experts suggested considering, which we summarize here. Too often the issues raised are seen to have technical fixes, or as discussed in the previous section, warranting participatory processes which may not adequately address the scope and scale of AI. We argue that the 3A Framework enables policymakers and practitioners to work through the issues holistically, and to identify relevant actors and responsibilities needed to include women in AI-enabled smart cities.

Returning to the two applications of AI for social good, CCTV and smart transportation systems integrate complex AI applications (Section 3). Socially good outcomes, especially for women, are not guaranteed. Developing and implementing these applications frequently involves multiple government, private sector, and community-based organizations, and they build on prior systems and infrastructures that are culturally and context-specific. Working towards socially good outcomes for diverse women, particularly the most marginalized, requires an effective distribution of roles and responsibilities across an interlaced network of actors.

The views expressed in Section 5 illustrate that from a big picture standpoint, a range of actors belonging to national government institutions, international organizations, inter-organizational coordination bodies, and workers' unions, etc. have a role to play in making conditions and opportunities more equitable for women across Asia-Pacific in the long-run. The experts clearly articulated that root causes, rather than symptoms of inequity, need to be addressed. However, viewed from such a vantage point, one might consider that different actors may attribute particular exclusion issues to different root causes. From an institutional perspective, this may be due to varied missions and objectives. Moreover, as Ruhiya Seward pointed out, institutions operate according to their own organizational logic and may have specific challenges to which they must attend. This suggests that national governments have a role to play in clarifying roles and responsibilities, especially in terms of the commitments they hold to gender equity. For instance, by creating stronger and more explicit connections in policy roadmaps between smart city plans and the achievement of the SDGs related to gender equality (e.g., Goals 5 and 11), making this information readily available and accessible is necessary.

In terms of the internal dynamics of AI-enabled systems, the most critical issue for the experts related to the need to expose the power relations at play. Both CCTV and smart transportation systems may be used for surveillance, and it is not clear what measures are in place to inform the public or take any of the unique concerns women hold for their safety and well-being to heart. The 3A Framework facilitates discussions surrounding power differentials to take shape, including a range of individual, community, and place-based aspects. A major impediment is that relationships between key decision makers of smart city initiatives and women are not well-established. Successful examples provided by the experts reflected how women, particularly the most marginalized, are more comfortable forming relationships within their communities, as in Anita Gurumurthy's example of community-driven water sanitation; or when there

is greater trust and transparency, as with Hannah Thinyane's example of the Apprise system for frontline workers (Section 5.1).

In parallel, private sector actors, such as those managing CCTV or transportation systems and intermediating between government and citizen groups, have an important facilitating role to play. These actors need to take time to understand local dynamics and ultimately help broaden and deepen the design, implementation, and management of AI technology in context, primarily by interacting with critical actors such as women's activist groups and community groups. It is only when system operators are aware of the interlaced network of actors and patterns of exclusion that they have the opportunity to use their power to encourage and provide entry points to systemic decision-making processes. Nevertheless, as Sue Keay reminds us, such actors are not likely to take on such responsibility unless these tasks are mandated and reported on. National and municipal governments must set high expectations of private sector actors to work more collaboratively with community groups. Sanctions could also be instituted as a means to hold industry partners accountable for more than delivering technologies and systems alone.

What Thinyane's research also demonstrates, however, is that AI technology may also assist in developing trusting, inclusive relationships if designed responsively and supported holistically. The 3A Framework does not discriminate between human or technological actors, or collections of these. There is scope for future work developing AI to find patterns of exclusion, to look for risks and breaches in a system, or to find patterns that are exclusive to marginalized women and which may assist in a greater proportion than other sections of a society. For example, by suggesting a public transport route or by optimizing routes when stops are permitted in between stops at night to enable women to disembark closer to their homes. Padmini Murray's work points to innovation in designing AI to build consensus in local governance, which may facilitate rebalancing the age-old power
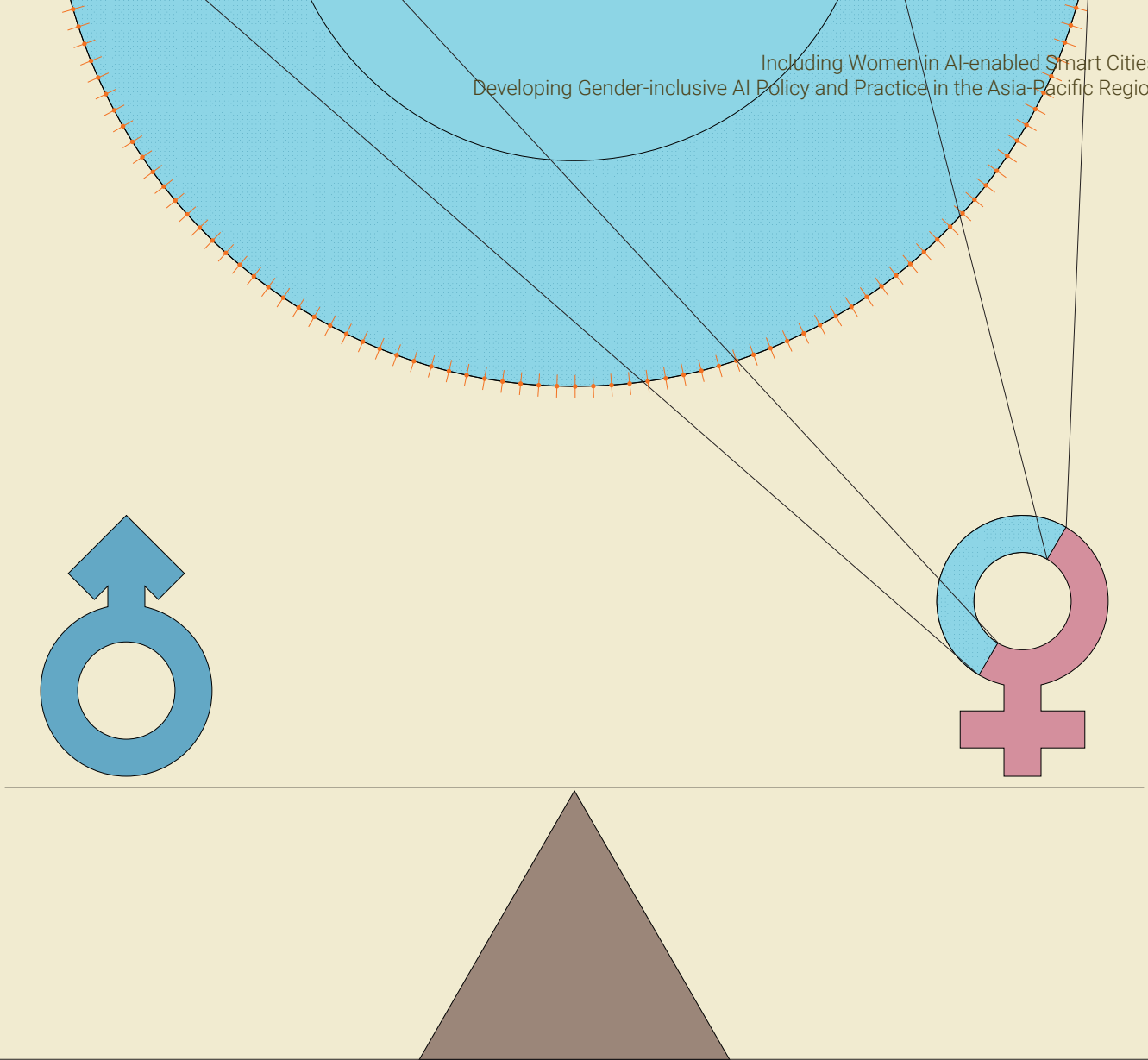
issues that have plagued participatory decision-making for diverse women. On the other hand, IT for Change has been exploring community ownership of data resources that has likewise improved inclusion outcomes. Progress in these areas suggests that AI, in terms of its design and features, will have a role and certain responsibilities in making AI-enabled smart cities more inclusive to women.

Further research is needed to identify the roles and responsibilities that will enable the holistic integration of both the big and more granular pictures in smart city developments. We argue that a new class of practitioners able to mobilize and circulate across the network of actors is needed. These practitioners will require a plethora of knowledge and skills to translate

between perspectives and make suggestions and improvements about how AI is designed, managed, and regulated in context. Another aspect identified by Anita Gurumurthy requiring further research is the influence of more powerful countries in Asia-Pacific on nations that have less capacity and resources to shape and control their own AI futures. Such intra-regional development may well impact on how the inclusion of women is taken up across the region (if, for instance, all countries begin to adopt the same AFRT system, and states are unable to modify or adapt it to their local context). Nevertheless, the 3A Framework may still be a useful tool for policymakers to use to navigate such tensions and global developments.

# References

Allwinkle, S., & Cruickshank, P. (2011). Creating Smart-er Cities: An Overview. *Journal of Urban Technology, 18*(2), 1–16. http://doi.org/10.1080/10630732.2011.601103

ASEAN. (2018). *ASEAN Smart Cities Network: Smart City Action Plans*. Singapore: ASEAN.

Australian Human Rights Commission. (2019). Human Rights and Technology Discussion Paper launches | Australian Human Rights Commission. Retrieved May 1, 2020, from https://www.humanrights.gov.au/about/news/human-rights-and-technology-discussion-paper-launches

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements. Psychological Science in the Public Interest, 20, 1–68. doi:10.1177/1529100619832930

Baruah, N. (2020). Her Right to the City: Must Women Tread in Fear? Retrieved May 1, 2020, from https://asiafoundation.org/2020/02/05/her-right-to-the-city-must-women-tread-in-fear/

Baud, I., Scott, D., Pfeffer, K., Sydenstricker-Neto, J., & Denis, E. (2014). Digital and spatial knowledge management in urban governance: Emerging issues in India, Brazil, South Africa, and Peru. *Habitat International, 44*, 501–509.

Baxter, W. (2017). Thailand 4.0 and the future of work in the Kingdom. Retrieved May 1, 2020, from https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/meetingdocument/wcms_549062.pdf

Bharucha, J., & Khatri, R. (2018). The sexual street harassment battle: perceptions of women in urban India. *The Journal of Adult Protection. 20*(2), 101-109. https://doi.org/10.1108/JAP-12-2017-0038

Bhavnani, K.-K., Foran, J., Kurian, P. A., & Munshi, D. (Eds.). (2016). Feminist Futures: Reimagining Women, Culture and Development (2nd ed.). London: Zed Books Ltd.

Cabinet Office, Government of Japan. (n.d.). Society 5.0. Retrieved May 1, 2020, from https://www8.cao.go.jp/cstp/english/society5_0/index.html

Cathelat, B. (2019). *Smart Cities: Shaping the Society of 2030*. Paris: UNESCO and NETEXPLO.

Chan, J. K.-S., & Anderson, S. (2015). *Rethinking Smart Cities—ICT for New-type Urbanization and Public Participation at the City and Community Level in China.* Beijing: Intel & UNDP.

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review, 43*(6). http://doi.org/10.2307/1229039

Daniell, K.A. (2012) *Co-engineering and participatory water management: organisational challenges for water governance*, UNESCO International Hydrology Series, Cambridge University Press, Cambridge, UK.

Davy, J. (2019). What lies ahead of Indonesia's 100 smart cities movement? Retrieved May 1, 2020, from https://www.thejakartapost.com/life/2019/12/05/what-lies-ahead-of-indonesias-100-smart-cities-movement.html

Department of the Prime Minister and Cabinet. (2016). *Smart Cities Plan*. Canberra: Commonwealth of Australia.

Empower Foundation. (2012). Hit & Run: Sex Worker's Research on Anti trafficking in Thailand. Bangkok: Empower Foundation.

Finlay, A. (Ed.). (2019). *Artificial Intelligence: Human Rights, Social Justice and Development*. New York: APC, Sida & Article 19.

Friedmann, J. (1992). Empowerment: The Politics of Alternative Development (1st ed.). Oxford: Wiley-Blackwell.

Ghazal, B., ElKhatib, K., Chahine, K., & Kherfan, M. (2016, April). Smart traffic light control system. In 2016 third international conference on electrical, electronics, computer engineering and their applications (EECEA) (pp. 140-145). IEEE.

Gekoski, A., Gray, J. M., Adler, J. R., & Horvath, M. A. (2017). *The prevalence and nature of sexual harassment and assault against women and girls on public transport: an international review.* Journal of Criminological Research, Policy and Practice, 3(1), 3-16, https://doi.org/10.1108/JCRPP-08-2016-0016

Giest, S. (2017). Big data analytics for mitigating carbon emissions in smart cities: Opportunities and challenges. *European Planning Studies, 25*(6), 941–957. http://doi.org/10.1080/09654313.2017.1294149

Global Slavery Index (2019). Asia and the Pacific | Global Slavery Index. Retrieved May 1, 2020, from **https://www.globalslaveryindex.org/2018/findings/regional-analysis/asia-and-the-pacific/**

Grother, P., Ngan, M., Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Washington, DC: U.S. Department of Commerce, National Institute of Standards and Technology.

Haque, M. M., Chin, H. C., & Debnath, A. K. (2013). Sustainable, safe, smart—three key elements of Singapore's evolving transport policies. Transport Policy, 27, 20-31.

Heise, L., Ellsberg, M., & Gottmoeller, M. (2002). A global overview of gender-based violence. *International Journal of Gynecology & Obstetrics,* 78, S5-S14.

Höjer, M., & Wangel, J. (2015). Smart sustainable cities: Definition and challenges. In L. Hilty, B. Aebischer (Eds.), *ICT innovations for sustainability* (Vol. 310, pp. 333–349). Cham: Springer, Cham.

Hörold, S., Mayas, C., & Krömker, H. (2015, August). Towards paperless mobility information in public transport. In International Conference on Human-Computer Interaction (pp. 340-349). Springer, Cham.

Hollands, R. G. (2008). Will the real smart city please stand up? *City, 12*(3), 303–320. **http://doi.org/10.1080/13604810802479126**

Huawei Enterprise. (2019). Smart City Framework and Guidance for Thailand: Smart City services for Phuket. Retrieved May 1, 2020 from **https://www.huawei.com/th/industry-insights/technology/smart-city-framework-and-guidance-for-thailand-smart-city-services-for-phuket**

Innovation and Technology Bureau. (2017). *Hong Kong Smart City Blueprint.* Hong Kong: Innovation and Technology Bureau.

Jackman, M. R. (2006). Gender, violence, and harassment. In B. Risman, C. Froyum, W.J. Scarborough (Eds.) *Handbook of the Sociology of Gender* (pp. 275-317). Boston: Springer.

Javaid, S., Sufian, A., Pervaiz, S., & Tanveer, M. (2018). Smart traffic management system using Internet of Things. In 2018 20th International Conference on Advanced Communication Technology (ICACT) (pp. 393-398). IEEE.

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal, 79*(1), 1–14. **http://doi.org/10.1007/s10708-013-9516-8**

Korea Legislation Research Institute. (2019). *Spatial Data Industry Promotion Act*. Retrieved May 1, 2020 from http://elaw.klri.re.kr/eng_service/lawView.do?hseq=38429&lang=ENG

Laksmi, S. (2018). *Buku 2: Master Plan Smart City Kabupaten Pati.* Retrieved May 1, 2020, from https://www.patikab.go.id/v2/uploaded/2019/Buku%202%20-%20MasterPlan%20Smart%20City%20Pati.pdf

Lam, P. T., & Yang, W. (2020). Factors influencing the consideration of Public-Private Partnerships (PPP) for smart city projects: Evidence from Hong Kong. *Cities, 99*, 102606, https://doi.org/10.1016/j.cities.2020.102606

Lee, J. J. (2005). Human trafficking in East Asia: current trends, data collection, and knowledge gaps. *International Migration, 43*(1-2), 165-201.

Long, Y., Zhang, E., Zhang, Y., Chen, Y., & Chen, Y. (2019). B*rief Review for Smart Cities of the Planet: Research Report* (pp. 1–23). Beijing: Hitachi China & Tsinghua University.

MacQueen, K. M., McLellan, E., Kay, K., & Milstein, B. (1998). Codebook development for team-based qualitative analysis. *Cultural Anthropology Methods, 10*(2), 31–36.

Marsal-Llacuna, M.-L. (2015). Building Universal Socio-cultural Indicators for Standardizing the Safeguarding of Citizens' Rights in Smart Cities. *Social Indicators Research, 130*(2), 563–579. http://doi.org/10.1007/s11205-015-1192-2

Martin, B. (2019). A Social Worker in the New Applied Science: An Individual Portfolio for CECS 6001, Fundamentals of a New Applied Science 1. Canberra: Australian National University.

Miles, M. B., & Huberman, A. M. (1994). Qualitative Data Analysis: An Expanded Sourcebook (2nd ed.). London: Sage.

Ministry of Communication and Information – Public Relations Bureau. (2017). *Tahap Pertama Gerakan Menuju 100 Smart City 2017, 24 Kota/Kabupaten Berhasil Menyelesaikan Smart City Masterplan*. Retrieved May 1, 2020, from https://www.kominfo.go.id/content/detail/11489/siaran-pers-no-223hmkominfo112017-tentang-tahap-pertama-gerakanmenuju-100-smart-city-2017-24-kotakabupaten-berhasil-menyelesaikan-smart-city-masterplan/0/siaran_pers

Ministry of Housing and Local Government. (2018). *Malaysia Smart City Framework*. Kuala Lumpur: Ministry of Housing and Local Government, Government of Malaysia.

Ministry of Housing and Urban Affairs. (2015). *Smart Pune: Creation of a Vision Community*. New Delhi: Government of India.

Ministry of Information and Communications. (2019a). Groundbreaking ceremony for first smart city project in Hanoi. Retrieved May 1, 2020, from https://english.mic.gov.vn/Pages/TinTuc/139806/Groundbreaking-ceremony-for-first-smart-city-project-in-Hanoi.html

Ministry of Information and Communications. (2019b). *Korean Smart Cities*. Seoul: Government of South Korea.

Ministry of Information and Communications. (n.d.). Smart city challenges Vietnamese Gov't and localities. Retrieved April 26, 2020, from https://english.mic.gov.vn/Pages/TinTuc/139821/Smart-city-challenges-Vietnamese-Gov-t-and-localities.html

Ministry of Urban Development (2015). *Citizen Consultations to Prepare Smart Cities Proposals (SCP)*. New Delhi: Government of India.

Ministry of Urban Development (2015). *Smart Cities − Mission Statement & Guidelines*. New Delhi: Government of India.

Musadat, A. (2019) Participatory Planning and Budgeting in Decentralised Indonesia: Understanding Participation, Responsiveness and Accountability, PhD Thesis, Australian National University, https://openresearch-repository.anu.edu.au/bitstream/1885/167001/1/Musadat2019PhDThesis.pdf

NITI Aayog. (2018). *National Strategy for Artificial Intelligence*. New Dehli: Government of India.

Oakley, P. (Ed.). (2001). Evaluating Empowerment: Reviewing the Concept and Practice (INTRAC NGO Management & Policy). Oxford: INTRAC.

Panori, A., Kakderi, C., & Tsarchopoulos, P. (2019). Designing the Ontology of a Smart City Application for Measuring Multidimensional Urban Poverty, 1−20. http://doi.org/10.1007/s13132-017-0504-y

Perkins, D. D., & Zimmerman, M. A. (1995). Empowerment theory, research, and application. *American Journal of Community Psychology, 23*(5), 569−579. http://doi.org/10.1007/BF02506982

Piper, N. (2005). A problem by a different name? A review of research on trafficking in South-East Asia and Oceania. *International migration, 43*(1-2), 203-233.

Plan International. (2016). *A Right to the Night: Australian Girls on their Safety in Public Spaces*. Sydney: Plan International Australia and Our Watch.

Planning and Urban Management Agency. (2013). *The Samoa National Urban Policy*. Apia: Government of Samoa.

Rao, T. (2017). Women's Safety Audit Walk Commences 16 Days. Retrieved May 1, 2020, from https://asiapacific.unwomen.org/en/news-and-events/stories/2017/11/womens-safety-audit-walk-commences-16-days

Roces, M. (2010). Asian feminisms: Women's movements from the Asian perspective. In M. Roces, & L. Edwards (Eds.), *Women's Movements in Asia: Feminisms and Transnational Activism*. London: Routledge.

Sadoway, D., & Shekhar, S. (2014). (Re) prioritizing citizens in smart cities governance: examples of smart citizenship from urban India. *The Journal of Community Informatics, 10*(3).

SAGE. (2018, May 23). Science in Australia Gender Equity (SAGE). Retrieved May 1, 2020, from https://www.sciencegenderequity.org.au/

SCC India Staff (2019). Financing smart cities in India. Retrieved May 1, 2020, from https://india.smartcitiescouncil.com/article/financing-smart-cities-india

Sey, A., & Hafkin, N. (Eds.). (2019). *Taking Stock: Data and Evidence on Gender Equality in Digital Access, Skills, and Leadership* (pp. 1–340). Tokyo: United Nations University.

Singh, Y. J. (2019). Is smart mobility also gender-smart? Journal of Gender Studies, http://doi.org/10.1080/09589236.2019.1650728

Smart Nation Singapore. (2018). *Digital Government Blueprint*. Singapore: Smart Nation Singapore.

Smart Nation Singapore. (2020). Pillars of smart nation. Retrieved May 1, 2020, from https://www.smartnation.gov.sg/why-Smart-Nation/pillars-of-smart-nation

Smart City Thailand. (2018). *Smart City Thailand: Annual Report 2018*. Bangkok: Smart City Thailand.

Taweesaengsakulthai, S., Laochankham, S., Kamnuansilpa, P., & Wongthanavasu, S. (2019). Thailand Smart Cities: What is the Path to Success?. *Asian Politics & Policy, 11*(1), 144-156.

Thinyane, H., & Bhat, K. S. (2019). Apprise (pp. 1–14). Presented at the 2019 CHI Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/3290605.3300385

Thrive (2018). How the private sector can partner on smart cities. Retrieved May 1, 2020, from https://thrive.dxc.technology/asia/2018/07/17/how-the-private-sector-can-partner-on-smart-cities/

Thynell, M. (2016). The quest for gender-sensitive and inclusive transport policies in growing Asian cities. Social Inclusion, 4(3), 72-82.

Transport New South Wales. (2020). The Challenge | Future Transport. Retrieved May 1, 2020, from https://future.transport.nsw.gov.au/technology/roadmap-in-delivery/transport-digital-accelerator/challenge

Trencher, G. (2019). Towards the smart city 2.0: Empirical evidence of using smartness as a tool for tackling social challenges. *Technological Forecasting and Social Change, 142*, 117–128. http://doi.org/10.1016/j.techfore.2018.07.033

UN DESA. (n.d.). Social Inclusion. Retrieved May 1, 2020, from https://www.un.org/development/desa/socialperspectiveondevelopment/issues/social-integration.html

UN Women. (2017). *Safe Cities and Safe Public Spaces: Global Results Report.* New York: UN Women.

UNESCO. (2019, February 22). Japan pushing ahead with Society 5.0 to overcome chronic social challenges. Retrieved May 1, 2020, from https://en.unesco.org/news/japan-pushing-ahead-society-50-overcome-chronic-social-challenges

Widadio, N. A. (2019). Many Indonesian women face sexually harassment: survey. Retrieved May 1, 2020, from https://www.aa.com.tr/en/asia-pacific/many-indonesian-women-face-sexually-harassment-survey/1658677

World Vision Australia. (2007). Human trafficking in Asia: Policy brief. Retrieved May 1, 2020, from https://www.worldvision.com.au/docs/default-source/publications/human-rights-and-trafficking/people-trafficking-in-the-asia-region.pdf

Zaugg, J. (2019). India is trying to build the world's largest facial recognition system. Retrieved May 1, 2020, from https://edition.cnn.com/2019/10/17/tech/india-facial-recognition-intl-hnk/index.html

Zhao, D., Dai, Y., & Zhang, Z. (2011). Computational intelligence in urban traffic signal control: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 485-494.

# Appendix 1: Technical specifications of AFRT and smart transportation systems

## AFRT: Overview of algorithms and factors influencing performance

Within AFRT systems, after the image(s) have been obtained for processing, they are transformed into a mathematical representation which is then compared with other representations of faces to obtain a similarity score. The similarity score is essentially the probability of a match between two faces (the sensed face and the previously recorded face). Various deep neural network methods have been developed to produce a similarity score. The vast majority of these use a labelled dataset to train the neural network to produce the correct result. In deployment, the neural network is then used to recognize patterns based on similar features found in previously classified images when compared to unseen images (Masi, 2018).

Therefore, these systems only perform well on pictures which are drawn from distributions similar to that of the training dataset. When software developers use biased or unrepresentative datasets to train an algorithm, error rates increase. This is especially problematic when AFRT systems are developed in foreign cultural contexts. For instance, for facial recognition systems developed in the US, the false match rate is the highest in East Asian populations, whereas for many (but not all) systems developed in East Asia, false positive matches between people born in East Asian countries are lower.

The demographics of the people included in the dataset is not the only factor which influences the generalizability of the dataset. If the images exhibit systematic biases, then these can also be learnt by the algorithm (and if they are also not present during implementation this will lead to error). For example, in the NIST report (Grother et al., 2019) underexposure of photographs of dark-skinned individuals was identified as a possible source of bias. The types of cameras used can mitigate possible sources of bias by providing more consistent images. For instance, verification systems that take images using infrared sensors provide more consistent illumination in different lighting conditions. Some commercial face verification algorithms (such as Apple's Face ID) instead use a depth image or are used in conjunction with a colored or monochrome image. Including depth information reduces false matches and makes it harder to spoof such systems by, for example, printing an image of a person's face.[2] Depth can also be used for identification systems, but getting accurate and high-resolution depth is harder when the person's face is far away from the sensor. However, it is important to make sure that the system is trained using the same type of images that will be used during deployment.

Another factor which can influence performance is the threshold, which determines how similar two images need to be before they are considered a match (according to the similarity score outlined earlier). The system is not likely to make perfect predictions, so trade-offs occur between the number of false and true matches[3] – suitable trade-offs depend on the application. Consider, for example, unlocking a phone

---

2. Including depth is not the only way to combat spoof attacks, see Ramachandra and Busch (2017) for an overview of spoof detection methods.
3. The Relative Operating Characteristic (ROC) curve is one way of examining the trade-offs for various thresholds for binary classification problems

using facial verification: setting a high threshold is feasible because the user can retry at different angles and a back-up method exists for unlocking, such as a pin. In contrast, if facial identification is used to search for trafficked women or perpetrators of violence, a lower threshold may be appropriate, especially if combined with a human review before intervention. Developers also try to improve performance by grouping images into demographics, which essentially sorts the images before they are analyzed. However, classifying individuals into demographics can be hurtful to people if they are misclassified[4] and the number of demographics which can be usefully defined is likely to be limited (e.g., by the availability of training data for each demographic), and so useful demographics may never be suitable for everyone.

## AFRT: Extensions

There is speculation about the other possible functionalities that could be built into AFRT systems to support public safety and security. An integrated CCTV system in Shenzhen, China has been designed to supposedly "formulate behavior prediction based on facial and behavioral reaction" (Huawei Enterprise, 2019, p. 74). However, a major review found no evidence that emotional states can be accurately inferred from the analysis of facial movements alone, without reference to culture or context (Barrett et al., 2019). There are also vision-based systems for detecting unusual behavior which have been proposed in academia (Xiang and Gong, 2008; Wiliem et al., 2012) and implemented in commercial products (Rhombus Systems, 2019). Behavioral prediction algorithms, which may help to identify struggles, health crises, or other aspects often use unsupervised learning techniques to detect "unusual" behavior, and

would still need human interpretation. There is also no evidence to suggest that the situations which women face are being factored into technological design and development of such systems.

## Smart transportation systems: objectives and constraints

The efficiency gains derived from smart traffic lights focus on optimizing traffic flows based on real-time monitoring of traffic conditions. Data on traffic conditions is collected using vehicle detection sensors, and is either used to determine optimal timing for a single traffic light or transmitted over the Internet to a data processing center where it is automatically analyzed to determine optimal traffic lights for a broader system. What "optimal" means will depend on how designers have encoded the priorities to optimize for into the system. For instance, there will be a trade-off between efficiency of the overall traffic (which has environmental implications) and incentives designers may want to introduce, such as prioritizing cyclists, public transport vehicles,[5] or emergency vehicles (Javaid, 2018; Ghazal et al., 2016). It is common for smart traffic lights to control a single traffic light without connection to a larger network, thus taking into account the volume of traffic to shorten or lengthen the amount of time a light remains green. As the system becomes more complex (e.g., controlling multiple lights, balancing multiple priorities, monitoring performance for a variety of well-travelled and less-travelled routes, and ensuring that people on less-travelled routes do not have to wait unreasonable amounts of time) more advanced algorithms and computational resources are required. This is the primary application of AI in this context. Zhao et al. (2012) found fuzzy logic, artificial neural

---

4. For example, gender detectors can be hurtful to members of the transgender community (Hamidi et al., 2018; Keyes, 2018).

5. Copenhagen is a good example of this – State of Green (2016); Rasmussen (2018); Copenhagen Technical and Environmental Administration (2011).

network, evolutionary and swarm, reinforcement learning and adaptive dynamic programming, and agent and game methods are common. Given the complexity and development time required to test and implement solutions to complex traffic flow management problems, it seems problematic that gendered preferences and perspectives have not been considered here.

Efficiency gains in smart public transport are envisaged in a similar manner. Public transportation services can be integrated within the same traffic management system to both prioritize public transport vehicles over private vehicles at intersections, as well as to inform route optimization to service popular routes effectively and avoid congestion. As such, smart traffic management systems usually include an end-to-end platform that users have access to (usually a mobile application). People can use the platform to plan, book, and pay for their journeys, as well as access real-time information about their transport (Hörold et al., 2015). The platform can include a range of transport options beyond traditional buses and trains, such as bike/car sharing or hire options. Singapore, for instance, has a system in place that manages its public trains and buses, and integrates private shuttle buses servicing social housing and condo blocks (Haque et al., 2013). Its payment system functions across these services and enables monitoring of journeys from start to finish. In these systems, data protection practices would need to be carefully designed and incorporated to comply with privacy laws and consumer expectations. In the case of smart public transportation, efficiency gains are built into existing systems and networks. If there are mobility issues that a woman experiences which are not addressed in the existing system, there does not

seem to be any specific functions or procedures in place to address them.

When considering potential smart traffic systems, a critical aspect is the underlying infrastructure requirements that affect both the traffic management system performance and how diverse women may benefit differently from it. For complex traffic management systems, it is crucial to have a strong and reliable Internet network for the sensors, control center, and traffic lights to be able to communicate in real-time and to be responsive to the current conditions. All smart traffic light systems depend fundamentally on a high density and dispersion of networked vehicle sensors to provide enough real-time data for meaningful decision-making. These may include microwave radar (Ho and Chung 2016), video (Javaid et al., 2018), motion sensors (e.g., using infrared transmitters and receivers) (Ghazal et al., 2016; Jagadeesh et al., 2015), and under road sensors – including induction loops and various weight in motion estimation systems, which can be based on technologies such as piezoelectric, capacitive mats, bending plates, load cells, and optical (Hancke and Hancke, 2013) – with some sensors focused on detecting pedestrian and cyclist traffic. Some work also suggests using smartphones as distributed sensors (Anagnostopoulos, 2016; Wang et al., 2012; Jayapal and Roy, 2016) although this usually relies on the cooperation of the smartphone owners and could disadvantage those who do not own or regularly carry a smartphone. Such extensive infrastructure and resource requirements have severe implications on the types of roads and neighborhoods in which these systems can be built. Women with the most need for mobility support may, in contrast, live in places where it is not possible to construct these systems.

# References

Anagnostopoulos, T., Ferreira, D., Samodelkin, A., Ahmed, M., & Kostakos, V. (2016). Cyclist-aware traffic lights through distributed smartphone sensing. *Pervasive and Mobile Computing, 31*, 22-36.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements. *Psychological Science in the Public Interest*, 20, 1–68. doi:10.1177/1529100619832930

Copenhagen Technical and Environmental Administration. (2011). Good, Better, Best: The City of Copenhagen's Bicycle Strategy 2011-2025. Retrieved May 1, 2020, from **https://www. eltis.org/sites/default/files/case-studies/documents/copenhagens_cycling_strategy.pdf**

Ghazal, B., ElKhatib, K., Chahine, K., & Kherfan, M. (2016, April). Smart traffic light control system. In 2016 third international conference on electrical, electronics, computer engineering and their applications (EECEA) (pp. 140-145). IEEE.

Grother, P., Ngan, M., Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Washington, DC: U.S. Department of Commerce, National Institute of Standards and Technology.

Haque, M. M., Chin, H. C., & Debnath, A. K. (2013). Sustainable, safe, smart—three key elements of Singapore's evolving transport policies. *Transport Policy*, 27, 20-31.

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1-13).

Hancke, G. P., & Hancke Jr, G. P. (2013). The role of advanced sensing in smart cities. *Sensors*, 13(1), 393-425.

Ho, T. J., & Chung, M. J. (2016). Information-aided smart schemes for vehicle flow detection enhancements of traffic microwave radar detectors. *Applied Sciences*, 6(7), 196.

Hörold, S., Mayas, C., & Krömker, H. (2015, August). Towards paperless mobility information in public transport. In International Conference on Human-Computer Interaction (pp. 340-349). Springer, Cham.

Huawei Enterprise (2019). Smart City Framework and Guidance for Thailand: Smart City services for Phuket. Retrieved May 1, 2020 from **https://www.huawei.com/th/industry-insights/technology/smart-city-framework-and-guidance-for-thailand-smart-city-services-for-phuket**

Javaid, S., Sufian, A., Pervaiz, S., & Tanveer, M. (2018). Smart traffic management system using Internet of Things. In 2018 20th International Conference on Advanced Communication Technology (ICACT) (pp. 393-398). IEEE.

Jayapal, C., & Roy, S. S. (2016). Road traffic congestion management using VANET. In 2016 International Conference on Advances in Human Machine Interaction (HMI) (pp. 1-7). IEEE.

Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1-22.

Masi, I., Wu, Y., Hassner, T., & Natarajan, P. (2018). Deep face recognition: A survey. In 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI) (pp. 471-478). IEEE.

Ramachandra, R., & Busch, C. (2017). Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 50(1), 1-37.

Rasmussen, S. (2018). Green mobility and traffic safety in Copenhagen. Retrieved May 1, 2020, from **https://transops.s3.amazonaws.com/uploaded_files/2018%20 AASHTO%20International%20Day%20-%20Panel%201%20-%20Rasmussen%20-%20 City%20of%20Copenhagen.pdf**

Rhombus Systems. (2019). Introducing Unusual Behavior Detection (UBD) – Human Stance, Behavior, and Fall Detection. Retrieved May 1, 2020, from **https://www.rhombussystems. com/blog/ai/introducing-unusual-behavior-detection-ubd-%E2%80%93-human-stance-behavior-and-fall-detection/**

State of Green. (2016). Sustainable Urban Transportation. Retrieved May 1, 2020, from h**ttps://stateofgreen.com/en/uploads/2016/06/Sustainable-Urban-Transportation.pdf**

Wang, W. Q., Zhang, X., Zhang, J., & Lim, H. B. (2012). Smart traffic cloud: An infrastructure for traffic applications. In 2012 IEEE 18th International Conference on Parallel and Distributed Systems (pp. 822-827). IEEE.

Wiliem, A., Madasu, V., Boles, W., & Yarlagadda, P. (2012). A suspicious behaviour detection using a context space model for smart surveillance systems. *Computer Vision and Image Understanding*, 116(2), 194-209.

Xiang, T., & Gong, S. (2008). Incremental and adaptive abnormal behaviour detection. *Computer Vision and Image Understanding*, 111(1), 59-73.

Zhao, D., Dai, Y., & Zhang, Z. (2011). Computational intelligence in urban traffic signal control: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), 485-494.

# AI and the Future of Work:

# A Policy Framework for Transforming Job Disruption into Social Good for All

**Wilson Wong**
Associate Professor,
Data Science and
Policy Studies Programme,
The Chinese University of Hong Kong

## Introduction: Policy as the Key to AI Promises

This paper examines the impact of artificial intelligence (AI) on the future of work to develop a policy framework for transforming job disruption caused by AI into social good for all. With the rapid advancement and progress of AI technology, there is little doubt that the era of AI will have an unprecedented impact on societies, economies, and governments in a significant and profound way with long-term effects and implications (Kaplan, 2016; OECD 2019a). Among them is the effect on the employment market through job disruptions. This can be referred to as the general process of replacing existing jobs by AI automation with the simultaneous potential of re-creating new opportunities and positions, which is the primary focus of this study.

Although the ideal state of AI is clearly desirable, and its promised returns to society are attractive and potentially enormous, it should never be taken for granted or assumed to be implemented effortlessly and automatically. The enabling factors in terms of good governance and sound policies are often less emphasized and frequently neglected in the current discussion. The cost of not paying serious attention to the issues and problems of job disruption can be too high to bear as it would mean the possibility of countries not being able to make a successful and smooth transition to the AI economy (Deming, 2017). In the absence of equity and fairness, even if an AI economy is achieved, the goals of AI for social good and using AI to empower all people can be severely compromised. Without smart and effective policies to meet the AI challenge of job disruption, the disadvantaged and high-risk members of society would be displaced by AI automation and face economic hardship and social marginalization.

A major goal of this paper is to set up a policy framework on the role of the government as well as the policy responses it should make in order to address the concerns and challenges brought by AI job disruption. According to Kai-Fu Lee, a world-renowned expert and venture capitalist of AI, the total disruption of patterns of work and employment would lead to an alarming estimate of 40% for current jobs lost to AI (Lee & Moon, 2019). His estimate is echoed by the statistics of Organization for Economic Co-operation and Development (OECD) (see Figure 1). The combined share of jobs at high risk of automation and significant risk of automation is higher than 40% for the average of OECD countries. Even for countries such as Norway and Finland, which face a relatively lower risk than the global standard, their share of jobs threatened by AI automation is still over 30%. At the higher end, countries such as Greece, Turkey, Lithuania, and Slovakia are around 60%. Shockingly, even for countries such as Japan, which is an advanced economy, its share of jobs at risk is still more than 50%, meaning that one out of two members of the labor force would be affected by AI automation.

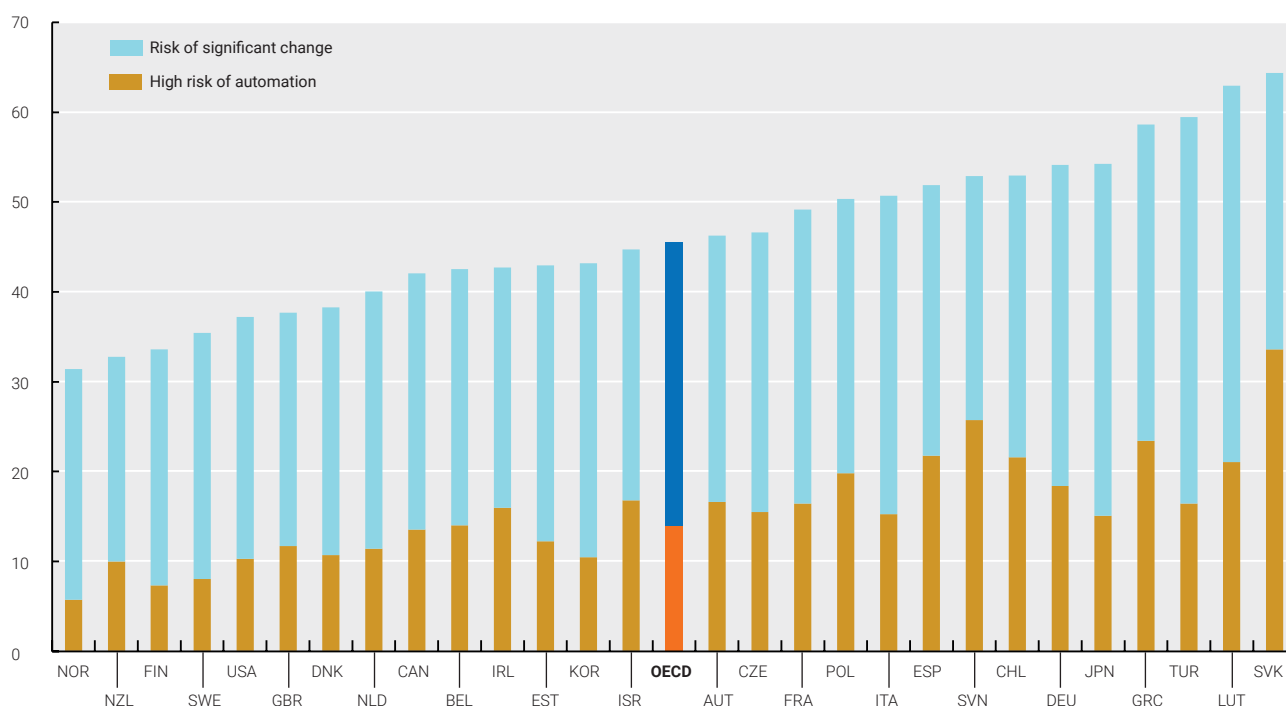## Large shares of jobs are at risk of automation or significant change



**Figure 1: Jobs at risk of automation in OECD countries**
*(Source: OECD The Future of Work 2019)*

In theory, with the widespread deployment of AI, nations and societies should win in the long run due to efficiency and productivity gains (OECD, 2018). However, with so much employment at risk under AI, in the short run, it is increasingly inevitable that there could be losers, of which include countries and citizens who are ill-prepared for the impact of AI. AI should be capable of creating a win-win outcome for all members of society (Lee & Moon, 2019). Any trade-off between labor rights and automation as well as tension between winners and losers should be a false dilemma. The key is whether proper policies are formulated and implemented to ensure all members of society can capture the benefits of AI.

As seen in Figure 2, the OECD Report of The Future of Work (2019b) finds that six out of ten adults lack the ICT skills necessary for the emerging jobs generated

by AI. Another alarming finding in the same report is that the most vulnerable population, whose jobs are at high risk under AI, are not being offered re-training or re-skilling opportunities. For example, for adults whose jobs face a high risk of automation, less than 20% of them are receiving re-training. Ironically, to the contrary, for adults whose jobs face low automation risk, close to 70% of them are receiving re-training. Similarly, less than 20% of low-skilled adults are receiving re-training, whereas up to 70% of high-skilled adults are undergoing re-training. All of these figures and statistics clearly show that there is a mismatch of policy and mistargeting in the allocation of resources for countries and governments in their transition to AI. Unless proper government policies are implemented in time, the future of AI could mean more inequalities in society and across nations, with the ambition of the technology unfulfilled.

### Skills and the future of work

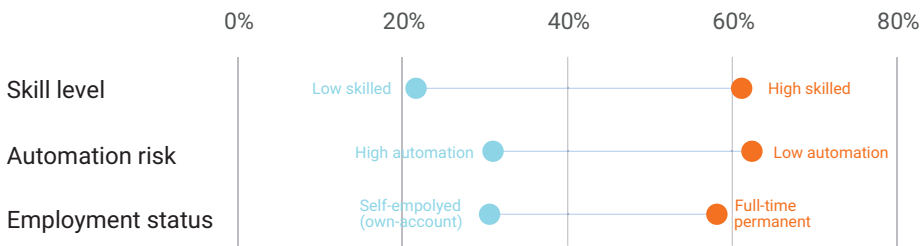Many adults do not have the right skills for emerging jobs

**6 out of 10 adults lack basic ICT skills** or have no computer experience..

**Share of highly-skilled jobs has increased by 25%** over last 2 decades. Low-skilled jobs have also increased, but middle-skilled jobs have decreased.

**25%**

Adults participation in training, by skill level, employment status and risk of automation.

| | 0% | 20% | 40% | 60% | 80% |
|---|---|---|---|---|---|
| Skill level | | Low skilled | | High skilled | |
| Automation risk | | High automation | | Low automation | |
| Employment status | | Self-employed (own-account) | | Full-time permanent | |

**Adult training should better target the disadvantaged.**

**TECHNOLOGY**

Can improve work-life balance (when, where and how to work)
Can create new opportunities
Tedious and dangerous tasks can be automated
Health and safety can be improved
Productivity boosted

**BUT LEARNING NEW SKILLS IS KEY**

**Figure 2: Skills and the future of work (OECD)**
*(Source: OECD Employment Outlook 2019: The Future of Work 2019)*

At the same time, a considerable gap has been observed between the demand for policy solutions and the supply of current knowledge on this topic. While there is a substantial amount of research and discussion on the impact of AI on economic growth and employment, there is relatively less research on what governments should do to turn the risk and threat of AI into job opportunities and social good for all. In the literature review conducted in this project, there is an evident shortage of relevant studies in the public policy and public administration literature to examine and analyze the proper role of governments and the required policy responses for addressing the impact of AI on the job market. Although this finding is concerning, many people believe that AI will have a major impact on the job market, including the issue of job losses and job elimination through automation. That said, there is limited knowledge on

what governments can do in order to address these adverse consequences (Kaplan, 2016). This is one of the key reasons why both policymakers and scholars must make greater efforts in preparing society for the AI era, especially because of its impact on policy, governance, and society (Desouza, 2018; Partnership for Public Service, 2018, 2019).

In bridging this gap, this paper will accomplish the following two major tasks: It first builds on the typology of job replacement and AI to set up a policy framework on the role of government and policy responses to address various concerns and challenges. On the principle of "rise with AI, not race with it" (World Bank, 2018), governments must play active or even aggressive roles not only on re-training, knowledge and skill building, and job re-creation, but also on social protection and a fair re-allocation of resources. Second, this paper conducts a survey of national AI strategies to assess the extent to which AI policy of job disruption is taken seriously by countries. It reveals that many countries, especially developing ones, are not well-prepared for AI, and most seem to be overlooking fairness and equity issues. In response, this paper suggests providing actionable policy recommendations to national governments and international authorities.

It is important to recognize that this paper is not an isolated effort in addressing these important questions and issues. Instead, it is a new step in a series of efforts by researchers and scholars of related projects to generate knowledge and findings substantiated by solid research on the social impact of AI and technology. More specifically, this is the second publication by the Association of Pacific Rim Universities (APRU) on technology and the transformation of work. It is hoped that this will build upon and extend the insights and findings of the first report, "Transformation of Work in Asia-Pacific in the 21st Century" published in 2019. This paper moves the collective project to the next stage by adopting a policy-oriented focus and a governmental approach to examine what governments should do to transform the threats and uncertainties of AI job disruption into opportunities for achieving social good for all.

# PART I: AI Impact on the Job Market

## The Typology of Job Replacement

Among the attempts to understand and theorize the impact of AI on the future of work, one of most useful and best-known frameworks for analyzing the effect of AI on the job market is the typology of job replacement developed by Lee Kai-Fu (2018) in his book "AI Super-Powers". His typology is shown in Figure 3. In basic terms, to analyze his framework, Lee uses two major dimensions: social nature of the job (social vs. non-social) and the degree to which the job can be replaced by automation (optimization-based vs. creativity or strategy based). Under this typology, four types of jobs with different effects under AI job replacement can be identified as below:
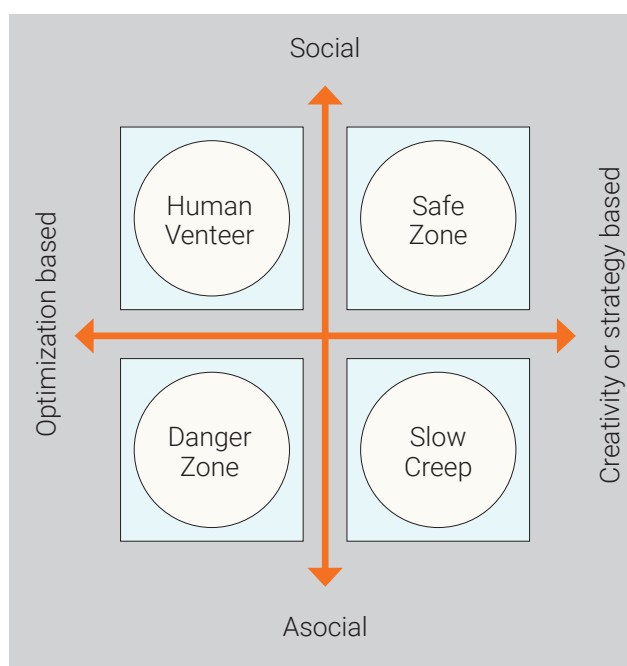


**Figure 3: A typology of risk of replacement by jobs**
*(Source: Lee (2019) and Author)*

### I. Danger Zone (non-social and optimization-based)
As evident by the title, jobs in the "Danger Zone" are those facing the highest risk of being replaced by AI automation (e.g., customer service representatives, drivers, basic translators, telemarketers, garment factory workers, chefs, and so on). These jobs face the most immediate danger of being replaced by

AI, and therefore should receive the highest policy priority. Low-skilled labor groups are often the most vulnerable, as they have limited access to re-training opportunities. Providing re-skilling and re-training to this group of people should create a win-win outcome. For society, higher efficiency can be yielded by eliminating "Danger Zone" jobs and replacing them with AI technology. For the workers concerned, through re-training and re-skilling, they can shift to job opportunities found in the other three quadrants, where they will experience higher productivity through taking advantage of AI, and as a result will enjoy higher wages.

## II. Human Veneer (social and optimization-based)

"Human Veneer" is a mixed and somewhat tricky category. In principle, most of the functions, tasks, and duties can already be done by AI, but the key social interactive element of the job makes it difficult to be fully automated (e.g., cafe waiters, wedding planners, teachers, doctors, hotel receptionists, and so on). If behind-the-scenes optimization work was completely taken over by AI, human actors would still be required as the social interface (the veneer) for clients and customers, representing the delicate performance balance and intricate symbolic relationship between AI and humans. This is exactly why bank tellers were not eliminated when the automated teller machine (ATM) was invented, as human interaction was still valued and preferred by many customers (Kang & Francisco, 2019).

According to Lee (2018), there are two factors which determine the percentage and how quickly jobs in the "Human Veneer" quadrant would be replaced by AI: the capability of restructuring the task and making AI more human-like in performing it; how open and receptive customers are to interacting with AI. Since the second factor can vary across cultures and social contexts, we can expect to see variations across countries on the type, degree, and pace of jobs being replaced by AI under "Human Veneer". In formulating proper policy response to job disruption, this quadrant underscores the importance of enhancing the social intelligence of workers in skill upgrade and re-training as it is a capability which cannot be performed and replaced by AI (OECD 2018).

## III. Slow Creep (non-social and creativity / strategy based)

The "Slow Creep" quadrant includes jobs which do not rely on human social skills, but would require another dimension of capacities which currently cannot be performed by AI: dexterity, strategic thinking, creativity, and the ability to adapt to an unstructured environment (OECD 2018; Frey and Osborne, 2017). Examples of jobs under this category include aerospace mechanics, scientists, artists, columnists, graphic designers, and security guards. This category is labelled as "Slow Creep" because it is generally believed that given the progress of AI technology and the advent of Big Data for AI training, it is plausible for AI to gradually narrow the gap with humans in terms of creativity and adaptation to uncertainties and contingencies. The pace of job elimination in this quadrant would depend less on process innovation in companies and organizations—a major factor affecting the job elimination in the "Human Veneer" quadrant—but would be more influenced by the progress and advancement of AI technology.

The special nature of "Slow Creep" has helped to accentuate the important principle advocated by the World Bank (2018) in the development of AI: "Rise with − not against − the Machine". In other words, humans should "rise with AI, not race with it" (World Bank, 2018). From a policy standpoint, it is pointless and fruitless for humans to have direct competition with AI, which is also contradictory to the intention of inventing new technology. Machines and technology are invented to aid humans—competing or replacing humans is not the objective. The development of AI should be human-centric for elevating the performance and strengthening the capacity of humans. Those in the "Slow Creep" category should be equipped with knowledge and skills of AI in order to enhance their ability to become more productive and creative.

## IV. Safe Zone (social and creativity / strategy based)

Jobs in the "Safe Zone" quadrant are those which possess two of the three major "engineering bottlenecks" (i.e., elements which cannot be easily automated by AI), such as social and creative

intelligence (Frey and Osborne, 2017). Some major examples of jobs under this category include CEOs, social workers, PR directors, dog trainers, physical therapists, and hair stylists. It is estimated that all of these jobs, due to their nature and the limitation of current AI capacities, are unlikely to be replaced by AI automation in the near and foreseeable future.

Nevertheless, it would be a mistake to take the "Safe Zone" as an "No-Action Zone" from a policy perspective. Job disruption policy should adopt a balanced, two-way approach to help those at a high risk of job replacement. This policy should also expand job opportunities and enhance the performance of people in the low-risk zone by upgrading their AI capacities. This should be the path leading to the overall goal of "AI for Social Good" and "AI for All" to benefit and empower all members of society. Although people with jobs in the "Safe Zone" quadrant face a much lower risk of losing their positions to AI, this does not exclude them from benefiting from AI itself. In this regard, workers and professionals in the "Safe Zone" should also be offered AI knowledge and skills through policy responses so that they can delegate more of their routine tasks to AI and fully concentrate on areas and duties in which they out-perform AI. In the meantime, many professionals

and staff in this quadrant are themselves leaders and changemakers in companies, governments, and non-profit organizations who can provide leadership and foresight in the development and adoption of AI in society through sectoral collaboration and other cooperative and engagement platforms.

## Job Disruption and the Generic Approach

Understanding the impact of job disruption should be a critical step towards formulating effective and appropriate policy responses. In this connection, some common misunderstandings and misperceptions about the effects of job disruption should be addressed here. First, job disruption impacts both physical and cognitive labor. All the above examples under each quadrant are taken from Lee's book "AI Super-Powers" (2018), which includes the jobs of both classifications. While there are debates and controversies about the suitability and correctness of each example, Lee's typology provides a useful framework for concretely and analytically understanding the effect of AI on the job market for providing a rigorous and scientifically based estimation of the effect on job loss, job elimination, and job disruption in the era of AI.
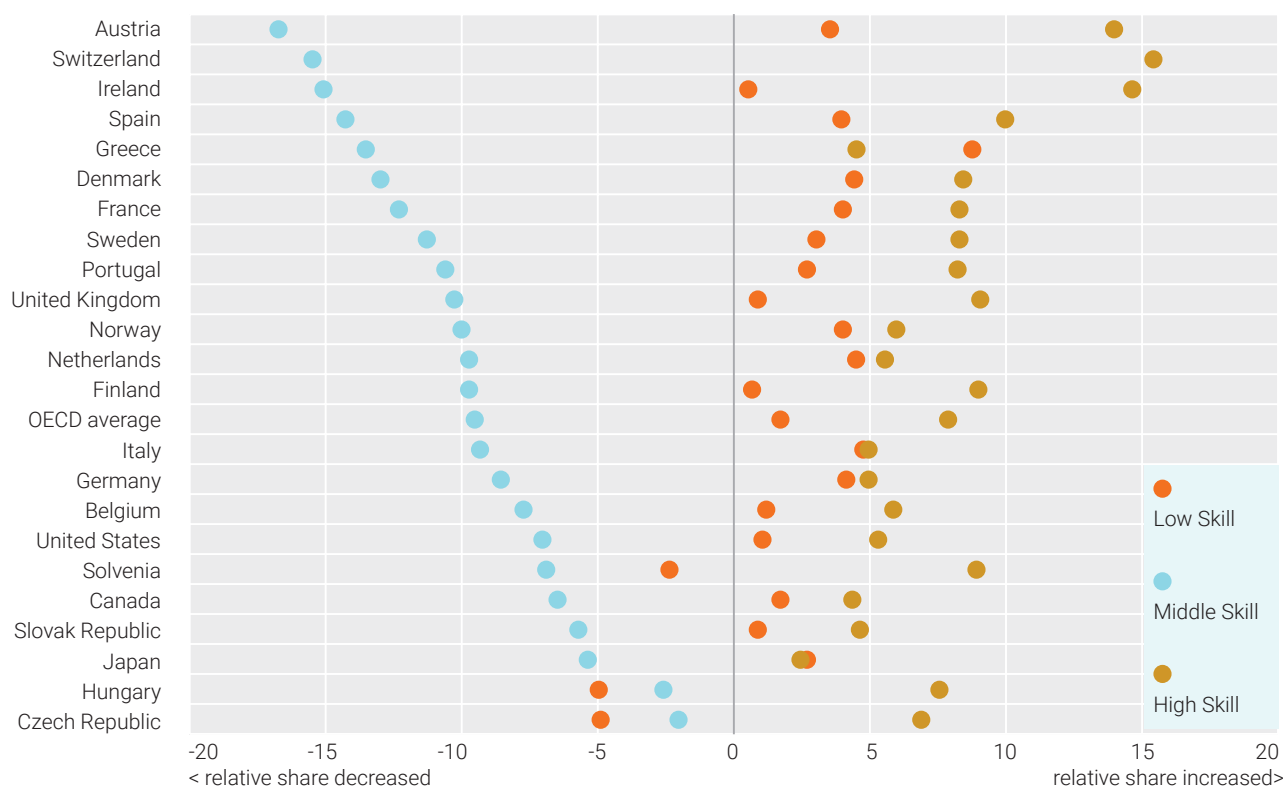


**Figure 4: Change of jobs by skill level (low, middle, high) in OECD countries (1995-2015)**
*(Source: OECD Employment Outlook 2017)*

A second common misperception is that AI automaton would only replace low-skilled jobs. Since cognitive labor is also at risk under AI automation, it is simply an oversimplification and is untrue. As shown by the examples provided in the above discussion, high-skilled and professional jobs such as teachers, doctors, and financial planners can still be replaced by AI. Skill level is not the most accurate and reliable indicator of whether a job would be disrupted by AI. It is still the two main factors: social intelligence and creative intelligence, which measure the risk of replacement. These two are limitations of the current technology of AI (Frey and Osborne, 2017), meaning that humans can keep their jobs as long as they can out-perform AI in terms of capacities and cost. To further substantiate this point (see Figure 4), between 1995 and 2015, middle-skill jobs were "disappearing", leading to a notable and intriguing situation of job polarization in the employment market in OECD countries. The average decrease of OECD countries in middle-skill jobs during this ten-year period was about negative 10%. In contrast, both low-skilled and high-skilled jobs have grown by about 2% and 7%, respectively.

The above numbers should be considered together with the change in manufacturing and non-manufacturing employment in OECD countries in the same period. Figure 5 shows significant shrinking of the manufacturing sector in many industries when there was remarkable growth elsewhere, such as the service industry. According to OECD (2019b), between 1995 and 2015, employment in the manufacturing sector declined by 20%, while increasing by 27% in the service sector. For example, employment in hotels and restaurants increased by over 40% and rose by about 20% in finance and insurance. After interpreting these figures, there are some key messages to take into consideration. First, most of the manufacturing jobs belong to the "Danger Zone" quadrant, which would explain their massive decline as a result of AI automation. Despite the fact that jobs are disappearing in this quadrant, new opportunities are being generated in other quadrants such as "Human Sheer" and "Safe Zone". This is why the non-manufacturing and service sectors are showing strong and robust growth, as many new jobs created belong to the other three quadrants.
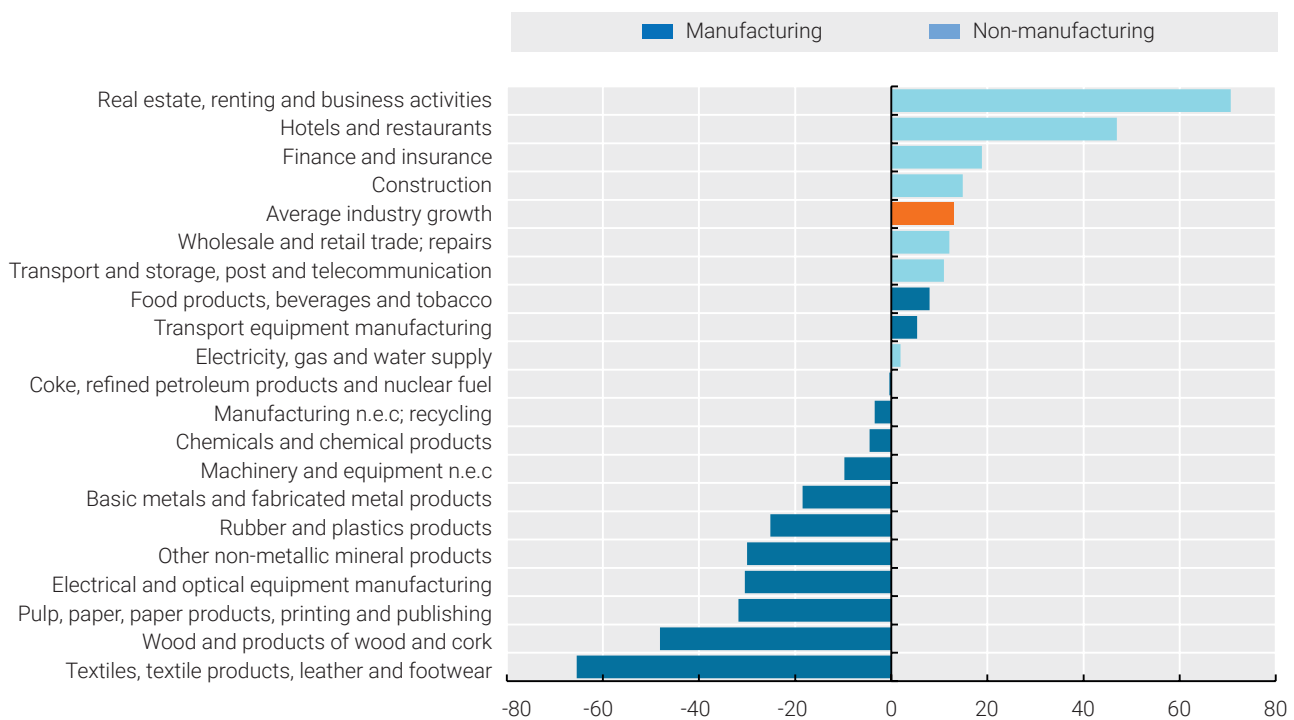


**Figure 5: The decline of the manufacturing sector in total employment within industry in OECD countries (1995-2015)**
*(Source: OECD The Furure of Work 2019)*

Recognizing a number of misperceptions, a more discerning and cautionary approach should be adopted in translating the findings of job disruption into policy implications. Even if employment can have a net and overall increase, there can be policy problems at both personal and country levels. For individuals, the government should offer re-skilling and re-training opportunities. At a country level, the government should invest heavily and strategically in AI infrastructure in order to build a labor force with AI knowledge and skills. When new job opportunities are created by AI, there is no guarantee that those jobs would necessarily be available in the country where the old positions were eliminated. New job opportunities pushed by AI can be created in advanced and developed countries, as poor and developing countries would likely suffer from huge job losses as a result of AI automation.

For this reason, without proper policies, AI automation can generate more inequities among individuals within society and internationally. In a free and global market, jobs and investment can move across national boundaries so that both individuals and countries can be AI-ready before receiving the benefits of AI (OECD, 2019b; World Economic Forum, 2018). This also reminds us of the importance and relevancy of context in assessing the impact of technology upgrades for any particular country (Kang & Francisco, 2019). For a country with poor AI infrastructure and low readiness of AI workforce, the rise of AI could potentially be devastating. This could cause a large-scale elimination of jobs as a result of AI, while new opportunities would be outflowed to other countries with a higher AI advantage.

While job loss and elimination under AI is inevitable, it can represent a "creative destruction" of the job market, as technology evolves and makes progress to create a brighter future for humankind (Schumpeter, 1942). Lee (2018) also gives a generally positive view of the future in which AI and humans can coexist in the labor market.

As shown in Figure 6, the only quadrant in which the co-existence of humans and AI is not possible is the "Danger Zone". However, in the other three quadrants, AI and humans can co-exist and reinforce each other in different modes and combinations in order to enhance performance and outcomes.
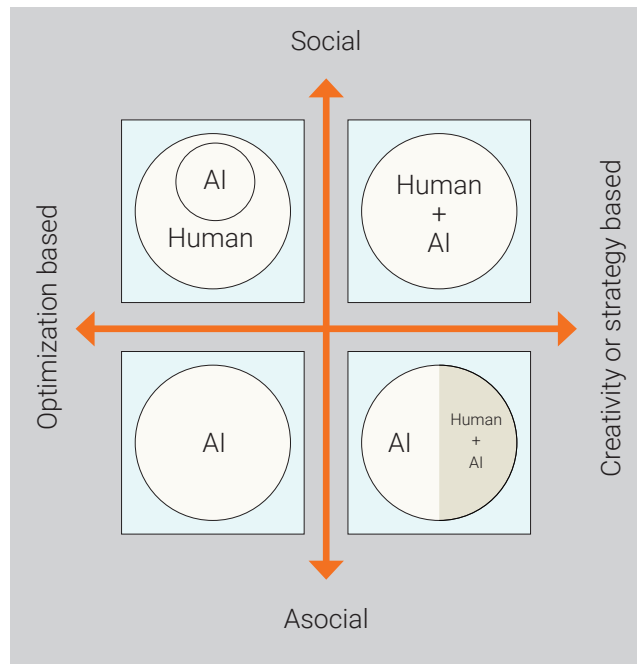


**Figure 6: Human – AI co-existence in the labor market**
*(Source: Lee (2018))*

To address the disruptive impact of AI on the job market, a generic "3R" approach has been developed: Reduce, Redistribute, and Retrain (Lee 2018). With regards to "Reduce," automation in the "Danger Zone" would reduce the working hours of many people. That said, people would work less but still enjoy the same standard of living. It is a symbol of the progress and prosperity of society in which AI provides more comfort and affluence. In principle, we can use redistribution, through means such as taxation and public expenditure, to shift resources from those who are still working (with higher performance) to those whose jobs have been replaced by AI. At the same

time, if there are still people who would like to stay in the job market, they can be "Retrained" (the third "R") to pick up the skills and knowledge required in the AI era (World Economic Forum, 2018).

The typology by Lee is consistent with and complementary to the other frameworks set up for evaluating the impact of AI on job disruption. Frey and Osborne (2017) have identified three types of tasks which cannot be easily replaced by AI and automation: perception and manipulation tasks, creative intelligence tasks, and social intelligence tasks. These three sets of tasks create serious challenges for codification and have been known as "engineering bottlenecks." Perception and manipulation tasks refer to tasks that are performed in unstructured, complex situations and handling irregular objects such as operating in cramped work spaces. "Creative intelligence tasks" refer to tasks that require original ideas. "Social intelligence tasks" need the understanding of other people's reactions in social contexts, or require assisting and caring for others.

Acemoglu and Autor (2011) have developed a helpful framework for assessing the impact of AI on wages and employment. Essentially, they divide technologies into two major types: enabling technologies and replacing technologies. Enabling technologies would help to expand the productivity of labor and therefore increase wages and job opportunities. Replacing technologies, such as manufacturing robots, would allow machines to be substituted for labor, which would result in jobs losses and wage reductions. From a standpoint of the whole society, both technologies are important to its progress. Yet, in formulating a labor and employment policy, the desirable direction should be to train an AI-competent labor force to work and rise with technologies to allow workers to benefit from the enabling technologies. This idea follows the guiding principle of "rise with AI, not race with it" (World Bank, 2018). Directing the labor force to compete with robots and AI in tasks related to replacing

technologies would only be a fatal, counter-productive, and irrational strategy (Kang and Francisco, 2017).

## The Policy Framework: Responses and Enabling Factors

When we consider the 3Rs in real-world settings, with real politics and policies, the situation would be much more complicated (Howlett & Ramesh, 1998; Kingdon, 1984; Lindblom, 2004). Many difficulties and obstacles would be encountered in addressing the impacts of technology such as AI on the job market in the complex and dynamic political environment (Ferro, et. al., 2013; Kitchin, 2014). For example, many people with jobs in the "Danger Zone" are believed to belong to the poor, older, and less educated segment of the population. For them, "retrain" and "redistribute" may not be preferable or politically feasible. Since they are old and less educated, re-training may not be realistic or affordable for them. In addition, poor people are often under-represented in politics, hence it would be unlikely for them to influence the government to have a re-distribution policy to compensate for job losses caused by AI and fund them for re-training programs. Resources used for redistribution must be generated from certain sources, such as those already benefiting from AI technology. However, companies and people profiting from AI are generally believed to be rich and powerful. It is therefore politically difficult to tax them in order to generate new resources to compensate those who would need help and assistance in adapting to the AI era.

To conclude, the 3Rs are an underestimation of the complexity and an oversimplification of the difficulties in the real-world policymaking process. Importantly, a well-developed and comprehensive framework does not exist, and therefore the 3Rs cannot be translated into effective and actionable policy responses. We also argue that 4Rs (i.e., "Rethink" as the fourth R) may be required in order to develop proper policy responses to address the challenges of AI.
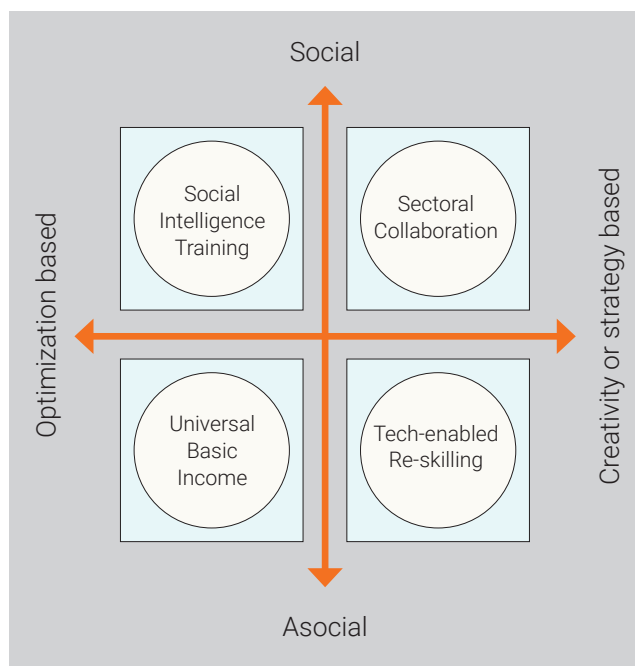
**Figure 7: Job disruption and policy responses**
*(Source: Lee (2019) and Author)*

Figure 7, Table 1, and Table 2 represent some of the initial but major efforts to set up a policy framework to address the policy issues and problems of AI on the job market. Figure 7 shows the major mode of policy response under each quadrant of job disruption. This does not exclude the possibility that there are many other complementary and compatible responses in each quadrant. However, the major mode represents the crux of the issues and concerns regarding the nature of the job category in the quadrant which should receive the most attention from policymakers.

Universal Basic Income (UBI) should be the major mode of policy when addressing the "Danger Zone". Even if these workers can take up new jobs in other quadrants after re-training and re-skilling, UBI should also be needed during the re-training period to support their lives and maintain their income. For the vulnerable population in the "Danger Zone", of which re-training and re-skilling would be less feasible due

to age, education, and other limiting factors such as health issues, UBI should become a long-term and stable source of income. In fact, this is closer to the original ideal of UBI in which all members of society should be unconditionally guaranteed a basic level of income, as AI should give rise to a rich society and provide a better quality of living for everyone.

Re-training is the key policy direction for both "Human Veneer" and "Slow Creep". That said, there is a subtle but important difference between the policy responses of the two. While the re-training in "Human Veneer" represents how to make humans more people-oriented, the re-training in "Slow Creep" should place more emphasis on enhancing the human capacity in mastering AI. This will enable them to be more creative and perform better at human functions and capacities that are unattainable by AI. In sum, re-training is more "human-oriented" in "Human Veneer" but should be more "technology-oriented" in "Slow Creep". It is untrue that the government has no role to play in the last quadrant of the "Safe Zone". To push AI technology forward and make sure it benefits future society, a partnership and collaboration among different sectors including governments, NGOs, universities, and industries, should be formed in order to lead the future development and application of AI technologies, rather than reacting to them passively.

Table 1 examines the impact of AI job disruption and policy responses by identifying the challenges and difficulties by type of job disruption and major policy mode. For example, it is expected that there would be significant problems regarding the politics of readjustment, transformation, and redistribution from vested interests after adopting innovative (but also controversial) policies such as UBI (Haggard, 1990; Polidano, 2001; Przeworski, & Limongi, 1993; Rodrik, 1992). Interest groups are powerful, and their rent-seeking activities often prevent the adoption of new technologies and slow down the progress and development of societies (Evans 1995; Johnson 1982; Kruger, 1974; Olson, 1982).

| Types of Disruption (ranked in terms of time urgency) | Policy Responses | Politics and Challenges |
|---|---|---|
| **Danger Zone (Reduce and Redistribute)** | • Universal Basic Income (UBI)<br>• Taxing AI and analysis of vulnerable population | • Politics of adjustment and transformation (sectoral vested interests)<br>• Politics of redistribution |
| **Human Veneer (Retrain)** | • Retraining and education (social intelligence)<br>• Life-long education (long-term education contract) | • Government partnerships with universities<br>• Reforming curriculum to eliminate the wall and divide between AI and human dimensions |
| **Slow Creep (Retrain)** | • Retraining and education (making humans more AI-equipped)<br>• Life-long education (long-term education contract) | • Reforming curriculum to eliminate the wall and divide between AI and human dimensions<br>• Government partnerships with universities |
| **Safe Zone (Rethink)** | • Exploring the opportunities, potential, and threats of AI<br>• Providing foresight and leadership | • Collaboration between multiple sectors (universities, governments, and industries)<br>• Balancing multiple and competing values in the process (including profit vs. social good) |

**Table 1: Policy and challenges in AI and job disruption**
*(Source: Author)*

As seen in Table 1, the changes required do not necessarily relate only to politics and institutional change; the change of role and mindset are equally as important. In this regard, universities play an irreplaceable role in leading AI technology and the creation of a knowledge-based learning society (Asia Development Bank, 2018; Florida, 2002). One of the major aspects, which requires a new mindset and fresh perspective, includes taking university education as a long-term contract between universities and citizens rather than a four-year commitment. "Students" are expected to return to campus much more frequently than before for training and education as new technologies arise. Besides, the wall and divide separating the boundary between human-centric liberal arts education and the technology-based STEM education should no longer be relevant and sensible in a world of AI. Critical revamping and radical restructuring of the curriculum in universities would be necessary to integrate the two into a single, coherent body of knowledge and skills to enable the new generation to be fully-equipped for the challenge and impact of AI (Tam, 2019; Yahya, 2019).

| Enabling Factors | Environment and Context |
|---|---|
| **Domestic level** | • Transparency and accountability in governance<br>• Participation and inclusive governance<br>• Fairness and justice in distribution and re-distribution<br>• Top-level government commitment<br>• Interagency task force<br>• Mechanisms for collaboration across sectors<br>• A knowledge-based learning society<br>• Active sector of university education<br>• Platform for learning and communication across universities, industries, society, and the government |
| **Human Veneer (Retrain)** | • A reliable and trustworthy international organization for learning and knowledge diffusion<br>• A regulation and enforcement framework on basic principles of AI<br>• International advice and support to eliminate the gap of "AI divide" between AI-rich and AI-poor countries |

**Table 2: Enabling factors – domestic and international levels**
*(Source: Author)*

Table 2 identifies the enabling factors for generating the policy responses in Table 1, and these factors are consistent with the major principles of good governance in the relevant studies and literature (Anderson, 2015; Cairney, 2016; Cath, 2018; Painter & Pierre, 2005). These factors can be divided into two major levels: domestic level and international level. At the domestic level, transparency, accountability, and participation should be some of the key elements in the public administration apparatus and decision-making process for formulating the effective and appropriate policy responses to AI job disruption. There should also be an inclusive and open process to ensure the involvement of all major stakeholders and actors in making all important policies. It would ascertain that the policy solutions are comprehensive and broadly supported for the welfare and benefit of all members of society, regardless of their political status and economic wealth. To facilitate the communication and collaboration of all actors and participants, a cross-sectoral platform should also be set up as the nexus of interaction and policymaking.

At the international level, organizations such as the United Nations (UN) and OECD should take the lead in major areas and capacities. Despite that, concrete and specific policy decisions should be conducted at the country level to respect its sovereignty while enabling it to design solutions that best fit its context (Welch & Wong, 1998). Despite this situation, international organizations and authorities can still make an outstanding and significant contribution to learning and knowledge diffusion by becoming a major hub of international AI cooperation (Straub, 2009). There should also be a key role for them to take up in establishing a regulatory framework on the basic principles of AI. If there is any area in which international organizations should have a more direct and close partnership with countries, it would be to provide resources and support to developing countries, which are most vulnerable to AI job disruption. Eliminating huge and detrimental international inequalities, the "AI divide" between AI-rich and AI-poor countries, should be a new and fundamental mission of international organizations in the AI era.

# PART II: Policy in Action – National AI Strategies

## The Survey

To assess the extent to which AI policy of job disruption is considered by major countries around the world, the second part of this paper conducts a survey of national AI strategies. The following major research and policy questions will be examined in this study. First, it attempts to find out if the impact of AI is being seriously considered at the country level, which can be easily reflected by whether or not the country has produced any open national document on AI strategy. If an AI strategy document exists, we would further examine its content and major initiatives, particularly the role of state and market in developing AI technology. In this regard, there are several possibilities and combinations: AI policy led by government, AI policy led by market, or AI policy led by a coalition of both government and market—a hybrid type of governance. Since the general goal of the market is profit-making, it is unlikely that a national AI strategy led mainly by it would be fair and equitable. With this in mind, we would also like to discover if equity and social protection are among the key areas emphasized in the national strategies. If so, what is the policy position and solutions that the country has raised in addressing these issues and concerns.

As an increasing number of countries prepare for the socio-economic transformation generated by AI, strategic documents are issued at various levels, crystallizing and encapsulating the vision and perspectives of top policymakers. A wide array of working group papers, consultations, guidelines, and reports precede and inform the design of a national strategy, but our analysis primarily focuses on the governmental strategies or national programs in their final form. These national AI strategy documents represent the policy consensus that are carefully-worded, influential, and committed. As a result, non-national AI strategy documents issued by non-state actors have not been included in this study. Preliminary, discussion, and consultation national documents on AI were also not selected, as they reflect more on "work-in-progress" or "initial thinking" than an adopted national policy position on AI. The

documents selected should also be dedicated to AI exclusively, as opposed to AI being listed together with other digital and ICT technologies. As the policy issue and concern is our center of attention, progress reports following up on national AI strategies have not been included in the study. These documents do not include new policy positions and mostly cover technical tools and the implementation details of these strategies. Furthermore, because the focus and scope of analysis of our study is national governments, documents issued by international organizations such as the UN, EU, and OECD have not been included. Despite this decision, the major content of relevant AI documents from these international organizations will still be summarized as a reference in the following sections.

This study follows a two-step methodological approach. First, starting from a comprehensive list of all national strategies complied from online research, all those which have an English version are selected according to our criteria stated above. The earliest AI national strategy document released is produced by South Korea, which can be dated back to as early as April 2016. The latest one included in the analysis is the National AI Strategy of Singapore, which was published in November 2019. After our selection, the national AI strategies will be analyzed in accordance with our research questions. A total of 15 documents by 12 countries were identified, collected, and analyzed (see Table 3). It should be noted that the actual number of documents would be much higher if some of our selection criteria was released. Because countries will continue to produce AI strategy documents, no list of such documents would be exhaustive. Since national AI strategy documents are a major policy communication tool for citizens, international partners, and stakeholders, we are confident that our study has included many important documents. They should also provide a representative sample of the state of AI national strategies for most countries throughout the world.

Using qualitative content analysis and comparative methods, the national AI strategies of Canada, China, Finland, France, Germany, Japan, Russia, Singapore, South Korea, Sweden, the United Kingdom, and the United States have been assessed to unveil their articulation of AI and its impact. This also includes their country-level policy responses to job disruption caused by AI automation. The analysis is driven by theoretical insights from the governance and ICT literature (Fountain, 2001; Norris, 2012; Wong, et. al., 2006), and thus should contribute to the current policy discussions by conceptually structuring the debates and offering a critical perspective of AI governance and the future of work.

| Date | Name of strategy | Country |
|---|---|---|
| **April 2016** | AI Information Industry Development Strategy | South Korea |
| **October 2016** | The National Artificial Intelligent Research and Development Strategic Plan | United States |
| **March 2017** | Pan-Canadian AI Strategy | Canada |
| **May 2017** | AI Program | Finland |
| **May 2017** | AI Technology Strategy | Japan |
| **July 2017** | Next Generation AI Development Plan | China |
| **March 2018** | AI Sector Deal | United Kingdom |
| **March 2018** | AI for Humanity | France |
| **May 2018** | National Approach to AI | Sweden |
| **November 2018** | Federal Government's AI Strategy | Germany |
| **February 2019** | Executive Order on Maintaining American Leadership in AI; and American AI Initiative | United States |
| **May 2019** | Beijing AI Principles | China |
| **June 2019** | The National Artificial Intelligent Research and Development Strategic Plans: 2019 Update | United States |
| **October 2019** | On the Development of AI in the Russian Federation | Russia |
| **November 2019** | National AI Strategy | Singapore |

**Table 3: National AI strategies included in the analysis**
*(Source: Author)*

## The AI Governance Landscape: Major Themes and Principles

AI has become a key focus of both national and international strategies, as their documents have been produced by individual countries and international organizations which are open to the public. For the latter, OECD published its "OECD Principles on AI" document in May 2019 and the EU released its "White Paper on Artificial Intelligence" in February 2020. Since international organizations generally have no jurisdiction over its member countries, their AI strategy documents tend to be guiding documents and a commitment to collaboration beyond state borders. They usually stand for agreement about continuing discussions on AI R&D and promoting cooperation to reach a human-centered AI society as well as reducing the risks of AI. They also include non-binding, principle-driven commitments that frame the international debate, highlighting the need to work together in order to remain competitive in AI.

The two recent documents by OECD and the EU provide excellent examples of the major points and observations above. In "OCED AI Principles", two of its five major principles are: "AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being", and "AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards—for example, enabling human intervention where necessary—to ensure a fair and just society". Similar statements, declarations, and principles have also been made by the EU. In the EU "White Paper on AI", shares and promotes the EU's vision of the benefits of AI to citizens, businesses, and public interest. For citizens, the EU believes that they should be able "to reap new benefits for example improved health care, fewer breakdowns of household machinery, safer and cleaner transport systems, better, and more accountable public services". In respect of public interest, the EU expects better and more efficient public services: "for services of public interest, for example by reducing the costs of providing services (transport, education, energy and waste management), by improving the sustainability of products and

by equipping law enforcement authorities with appropriate tools to ensure the security of citizens, with proper safeguards to respect their rights and freedoms."

National strategies, on the other hand, tend to be dominant, prescriptive approaches. They are unifying governmental documents that outline directions and priorities for domestic efforts and the allocation of resources. In some cases, they may apply to different levels of government in an uncoordinated manner, such as in the US. AI has generated an unprecedented number of national strategies and frameworks in a relatively short period of time. Although the field of AI can be dated back to the 1950s, the current development of AI strategy and regulation closely mirrors those of the Internet (Radu, 2019). This similarity can be linked to the fact that the Internet remains a key vehicle for "feeding" AI devices and for real-time experimentation with large amounts of data (Schonberger & Cukier, 2013).

It is not difficult to understand the background for the sudden surge of national AI strategies in recent years. Widely recognized as a disruptive technology (Bower & Christensen, 1995), AI is at the center of societal transformation, technology innovation, risk assessment, and governance debates. The ubiquity and extensive applications of AI corresponds with the focus of attention in AI discussions, which ranges from designing efficient systems and ensuring competitiveness to constructing ethical frameworks, risks assessment, legal responsibility, and certainly the impact on the human labor market and job disruption as AI advances.

With reference to the first research question in our study, the findings of our study are both striking and alarming. The number of countries which have national AI strategies (as defined by our selection criteria) are much fewer than expected—only 12 to be exact. In the UN membership, there are currently a total of 195 countries. This means that only 6% of them have a formal and well-articulated national AI

strategy to take advantage of AI and cope with its potentially negative impact; these figures cast doubt on their readiness for AI.

It is not only a small number of countries which causes concern here. The type of countries with or without a national AI strategy is worth discussing. A gap can also be found between the early adopters of AI strategies and countries which are still in the process of drafting a national policy. The first tend to be AI leaders and developed countries (e.g., the US, Germany, Japan, and South Korea), rather than developing countries (e.g., Laos, Nepal, Nigeria, and Myanmar). This validates and confirms the existence of an "AI divide" on a global scale. Out of the 12 countries in our survey, only China may still be considered as a developing country. However, this nation is clearly an exception rather than the norm given its economic power and international influence. A closer look at China would reveal that it is not a developing country from a typical sense, as it has attained the standard of many developed countries in terms of many major aspects, such as research and technology, and is a rising global power.

Since AI would impact both developed and developing countries, the poor preparation and low readiness of developing countries for AI automation should be a priority for the global policy agenda. Without proper policy responses at both country and international levels, it can be predicted that there would be a global AI-divide between developed and developing countries. There is a "race to the top" among AI-rich countries, but a "race to the bottom" among AI-poor countries. These two concurrent and parallel global races will eventually converge and quickly degenerate into enormous economic and social inequalities across countries. Similar gaps, such as the digital divide, have been observed from the differences in rates of progress, diffusion, and adoption of new technologies (Ake, 2001; Wong & Welch, 2004; Welch, Hinnant & Moon, 2005). They essentially reflect the contextual and institutional factors of the countries rather than the technical content and nature of the technology itself (Haque, 1996; Fountain, 2001; North, 1990; Painter & Pierre, 2005; Pollitt & Bouckaert, 2011; Wong, 2013).

The need for international cooperation is recognized by the majority of countries. Among EU member states, there is coherence around the perceived regional influence and work conducted at the supra-national level. Surprisingly, the relationship with developing countries is rarely mentioned. One exception is Germany, whose national strategy "Federal Government's Artificial Intelligence Strategy" has an action point to build up capacities and knowledge about AI in developing countries to promote economic cooperation and utilize economic and social opportunities.

**Highlights of major principles and objectives in the national strategies**

| Country | Major Principles and Objectives |
|---|---|
| **South Korea** | • Foster an intelligent information society on the basis of public-private partnership, with businesses and citizens playing leading roles and the government and research community providing support.<br>• Devise and implement a balanced policy regime that encompasses technologies, industries, and society and shapes the development of a more humane society.<br>• Provide strategic support for the prompt securement of the rights and access to Intelligent IT and other related resources to ensure and foster industrial competitiveness in advance.<br>• Reform policies and expand the social security net on the basis of social consensuses. |
| **United States** | Strategy 1: Make long-term investments in AI research<br>Strategy 2: Develop effective methods for human-AI collaboration<br>Strategy 3: Understand and address the ethical, legal, and societal implications of AI<br>Strategy 4: Ensure the safety and security of AI systems<br>Strategy 5: Develop shared public datasets and environments for AI training and testing<br>Strategy 6: Measure and evaluate AI technologies through benchmarks and standards<br>Strategy 7: Better understand the national AI R&D workforce needs<br>Strategy 8: Expand public-private partnerships in AI to accelerate advances in AI |
| **Canada** | The strategy has five major goals:<br>• Build a critical mass of talent within existing geographic areas of research excellence<br>• Increase the number of outstanding faculty in deep AI nationwide<br>• Dramatically increase the number of Canadian graduate and undergraduate students being trained in deep AI<br>• Create national programs that build a pan-Canadian AI community<br>• Position Canada as scientific leaders in AI research, and build on this science to ensure continuing prosperity and progress for all Canadians |
| **Finland** | Eleven key actions:<br>1. Enhance business competitiveness through the use of AI<br>2. Effectively utilize data in all sectors<br>3. Ensure that AI can be adopted more quickly and easily<br>4. Ensure top-level expertise and attract top experts<br>5. Make bold decisions and investments<br>6. Build the world's best public services<br>7. Establish new models for collaboration<br>8. Make Finland a frontrunner in the age of AI<br>9. Prepare for AI to change the nature of work<br>10. Steer AI development into a trust-based, human-centric direction<br>11. Prepare for security challenges |

**Table 4: Highlights of major principles and objectives in the national strategies**

| | |
|---|---|
| **Japan** | Basic Philosophies:<br>• Human-centered society<br>• Share guidelines as non-binding soft law with stakeholders internationally<br>• Ensure balance of benefits and risks<br>• Avoid hindering technologies or imposing excessive burdens on developers<br><br>9 Principles:<br>• Principle of collaboration   • Principle of transparency<br>• Principle of controllability   • Principle of safety<br>• Principle of security   • Principle of privacy<br>• Principle of user assistance   • Principle of accountability<br>• Principle of ethics (respect human dignity and individual autonomy) |
| **China** | Beijing AI principles:<br>• The R&D of AI should observe the following principles:<br>  do good; for humanity; be responsible; control risks; be ethical; be diverse and inclusive; open and share<br>• The use of AI should observe the following principles:<br>  use wisely and properly; informed-consent; education and training<br>• The governance of AI should observe the following principles:<br>  optimizing employment; harmony and cooperation: adaptation and moderation; subdivision and implementation; long-term planning |
| **United Kingdom** | Five Foundations<br>• Ideas - the world's most innovative economy<br>• People - good jobs and greater earning power for all<br>• Infrastructure - a major upgrade to the UK's infrastructure<br>• Business environment - the best place to start and grow a business<br>• Places - prosperous communities across the UK<br><br>Four Grand Challenges<br>• AI and Data Economy - We will put the UK at the forefront of the AI and data revolution<br>• Future of Mobility - We will become a world leader in the way people, goods and services move<br>• Clean Growth - We will maximize the advantages for UK industry from the global shift to clean growth<br>• Ageing Society - We will harness the power of innovation to help meet the needs of an ageing society |
| **France** | Primary themes:<br>1. Developing an aggressive data policy [to improve access to big data];<br>2. Targeting four strategic sectors [healthcare, environment, transport, and defense];<br>3. Boosting the potential of French research [and investing in talent];<br>4. Planning for the impact of AI on labor;<br>5. Making AI more environmentally friendly;<br>6. Opening up the black boxes of AI; and<br>7. Ensuring that AI supports inclusivity and diversity. |

**(Cont.) Table 4: Highlights of major principles and objectives in the national strategies**

| | |
|---|---|
| **Sweden** | The government's goals are to develop standards and principles – while acknowledging existing national and international regulations and norms – for ethical, sustainable, and safe AI; to continue to improve digital infrastructure to leverage opportunities in AI; to increase access to data; and to play an active role in the EU's digitization efforts. |
| **Germany** | The strategy pursues the following three objectives:<br>1. Making Germany and Europe global leaders on the development and use of AI technologies and securing Germany's competitiveness in the future;<br>2. Safeguarding the responsible development and use of AI which serves the good of society; and<br>3. Integrating AI in society in ethical, legal, cultural, and institutional terms in the context of a broad societal dialogue and active political measures. |
| **Russia** | Basic Principles of the Development and Use of AI Technologies:<br>a) The protection of human rights and liberties<br>b) Security<br>c) Transparency<br>d) Technological sovereignty<br>e) Innovation cycle integrity<br>f) Reasonable thrift<br>g) Support for competition |
| **Singapore** | This strategy serves three purposes:<br>1. Identify areas to focus attention and resources at a national level.<br>2. Set out how governments, companies, and researchers can work together to realize the positive impact of AI.<br>3. Address areas where attention is needed to manage change and/or manage new forms of risks that arise when AI becomes more pervasive.<br><br>Vision:<br>By 2030, Singapore will be a leader in developing and deploying scalable, impactful AI solutions, in key sectors of high value and relevance to our citizens and businesses (Smart Nation).<br><br>Approach:<br>1. Emphasize deployment<br>2. Focus on key sectors<br>3. Strengthen the AI Deployment Loop<br>4. Adopt a human-centric approach |

**(Cont.) Table 4: Highlights of major principles and objectives in the national strategies**

The major content of exemplary national AI documents is summarized in Table 4. The strategies analyzed here vary in scope and length, ranging from visions of development in the sector to full-fledged industrial strategies or comprehensive, all-sector approaches. Withstanding these minor differences, in general, there is a strong market orientation as the private sector traditionally takes the lead in AI research and development. For example, one of the major AI strategies of the US is to "expand public-private partnerships in AI to accelerate advances in AI". There is also an overwhelming and implicit assumption underlying all of these documents, which is the ability of AI to generate net positive social benefits. They also focus mostly on economic growth, national competitiveness and research, and investment. For the US, their number one strategy is to make long-term investments in AI research. In the same vein, Canada's top goal is to "build a critical mass of talent within existing geographic areas of research excellence." In the UK, a key national strategy for AI is transforming itself into "the world's most innovative economy." Unfortunately, equity and social protection is clearly not a significant topic in these national AI strategies, which seems quite alarming.

Global politics and international competition are major factors driving the increase in national AI strategies. The "global AI race" is often linked to the "great powers" discourse, which includes countries such as the US, Russia, and China, who are constantly competing for global dominance and supremacy (Lee, 2018). Apart from prevailing global powers, other major countries are eager to join the AI race. There is often a co-existence of a dual image in the documents—technical and political. On one hand, AI is presented in technical languages as relying on neutral networks modelling to mathematically analyze huge amounts of data for scientific and industrial revolutions. Politically, AI

development is considered crucial for the new race to the top among powerful nations. For instance, Canada would like to position itself as "a scientific leader in artificial intelligence research, and build on this science to ensure continuing prosperity and progress for all Canadians." For Germany, using its AI strategy, it pursues the objective of "making Germany and Europe global leaders on the development and use of AI technologies and securing Germany's competitiveness in the future." For France, one of its primary themes of AI strategy is "developing an aggressive data policy to improve access to big data." For the UK, its government aims to put the country "at the forefront of the artificial intelligence and data revolution."

National strategies are the first crucial step towards setting up a policy direction for AI. Their construction, articulation, production, and presentation in the public domain is a powerful political statement and a demonstration of national pride and supremacy. All major countries have the ambition of becoming the world leaders of this technology. Furthermore, some countries, such as China, have even taken a further step by highlighting their intention to drive the global governance of AI. From a historical perspective, the centrality of the nation state in AI debates is rather new (Radu, 2019). While international relations scholars have long reflected on the networked aspect of governance, where the state can be an orchestrator or partner, AI discourse at the national level brings forward a new dimension of state involvement in emerging technology regulation, which is in line with recent efforts to command control over strategic areas.

## The Missing Piece: Equity and Social Protection

Whilst national AI strategies focus primarily on economic growth, national competitiveness, and research and development, equity and social

protection is an important missing piece. It has never been a major topic or focus in the national AI strategy documents surveyed, and in some cases, it was simply ignored or forgotten. For example, no country has raised the idea of UBI, and social policy and re-distribution was not a key topic in any of the national AI strategy documents reviewed. Overall, social policy and readjustment in welfare programs do not seem to be the main concern. The job disruption problem is generally understood and framed as a re-training problem. It is also assumed that if significant wealth can be generated from AI development, there would be sufficient resources to handle other problems generated subsequently and naturally.

The mode and role of national AI strategy should be one of the main elements to reinforce the negligence and inattentiveness of equity and social protection for AI automation and job disruption. In most countries, hybrid governance—an alliance between government and market—is the primary driver of AI strategy for economic growth and national competitiveness. As reflected by the priorities of national AI strategy, the most urgent and primary concern of most countries is joining the private sector in the AI race to avoid being overtaken by other countries. With the market as the major partner, it is unlikely that a national AI strategy would result in a fair and equitable society.

In effect, AI governance is highly dominated by corporate interests. Overall, AI R&D continues to be driven by multinationals with headquarters concentrated in a few countries, while policy directions appear to be more reactive than anticipatory. From the patenting behavior of the largest companies between 2012 and 2014, the overwhelming majority (93%) of

AI patents were registered in Japan (33%), Republic of Korea (20%), USA (18%), Taiwan, China (8%), Germany (3%), and France (2%) (UNESCO, 2014). In international patent applications, China came second after the US last year (WIPO, 2018). A few companies from these two countries also have the largest AI research investments and development of standards, which has been further integrated in their products and services.

In consistency with the above trend, it is also common for the state to work alongside companies for financial investments in R&D. For example, the Canadian strategy focuses exclusively on research leadership and points to the use of government investment as a catalyst for investments from other levels of government and from the private sector. Following a similar approach, the UK and Germany mentioned export support for innovative AI and data businesses, as well as specific programs to attract such companies to establish headquarters on their territory, in addition to the use of trade missions abroad for their promotion. Moreover, the continuing interest and involvement of private actors is visible in the composition of oversight bodies or organizations driving the AI policy mandates, while nonprofit organizations and rights groups tend not to be equally well-represented (e.g., the UK and Canada).

Under the heavy influence of market ideology and the chief orientation on economic growth and national competitiveness, there is a strong tendency of using re-training and education in lieu of social policy and re-distribution in national AI strategies. The derived lack of serious concern and in-depth discussion on AI job disruption and the related remedial policies can be seen by examples shown in Table 5.

**Examples of explicit wording regarding education and social protection (if any) in national strategies**

| Country | Examples |
|---|---|
| **South Korea** | "Policy objective: Reform and tailor education, employment, and welfare services in response to changes in order to ensure that all citizens are able to enjoy the benefits of the intelligent information society."<br><br>"Foster and educate active workers capable of leading the intelligent information society based on their creativity and emotional intelligence. Ensure opportunities for a decent and humane standard of living by supporting the re-training of personnel and improving the employment and welfare environments." |
| **United States** | "Attaining the needed AI R&D advances outlined in this strategy will require a sufficient AI R&D workforce. Nations with the strongest presence in AI R&D will establish leading positions in the automation of the future. They will become the frontrunners in competencies like algorithm creation and development; capability demonstration; and commercialization. Developing technical expertise will provide the basis for these advancements." (*The National Artificial Intelligent Research and Development Strategic Plan, 2016*)<br><br>"The American AI Initiative is accelerating our Nation's leadership in AI. By driving technological breakthroughs in AI, breaking barriers to AI innovation, preparing our workforce for the jobs of the future, and protecting America's advantage in AI we are ensuring that AI technologies continue to improve the lives of our people, create jobs, reflect our Nation's values, and keep Americans safe at home and abroad." (*The American AI Initiative, 2019*)<br><br>"The United States must train current and future generations of American workers with the skills to develop and apply AI technologies to prepare them for today's economy and jobs of the future." (*Executive Order on Maintaining American Leadership in Artificial Intelligence, 2019*) |
| **Finland** | "The prerequisite for the broad-based utilization of artificial intelligence is that the population for the most part has a command of the skills and knowledge needed for its application. The requirements for the age of artificial intelligence should be visible in study content throughout the entire education system. At the moment, it is believed that the importance of skills related to social intelligence will grow.<br><br>The social security system must function flawlessly as people's working careers become diversified. Transitions between paid labor and entrepreneurship should be more flexible. Earnings level insurances misfortune allows for risk-taking in the broad sense. On the other hand, comprehensive earnings security insurance inevitably involves incentive problems. The long-term objective should be to increase the inventiveness of both social and unemployment security and improve the strengths related to these." |

**Table 5: Examples of explicit wording regarding education and social protection (if any) in national strategies**

| China | Vigorously strengthen training for the labor force working in AI. Accelerate the study of how AI affects the employment structure, the change of employment methods and the skills demands of new occupations and jobs. Establish lifelong learning and employment training systems to meet the needs of intelligent economy and intelligent society, and support institutions of higher learning, vocational schools and socialization training Institutions to carry out AI skills training, substantially increasing the professional skills of workers to meet the demands of the high-quality jobs in China's AI research. Encourage enterprises and organizations to provide AI skills training for employees. Strengthen re-employment training and guidance for workers to ensure that simple and repetitive work and the smooth transition of workers due to AI." (*Next Generation AI Development Plan, 2017*)<br><br>"Optimizing Employment: An inclusive attitude should be taken towards the potential impact of AI on human employment. A cautious attitude should be taken towards the promotion of AI applications that may have huge impacts on human employment. Explorations on Human-AI coordination and new forms of work that would give full play to human advantages and characteristics should be encouraged." (*Beijing AI Principles, 2019*) |
|---|---|
| United Kingdom | "People<br><br>• Establish a technical education system that rivals the best in the world to stand alongside our world-class higher education system<br>• Invest an additional £406 million in mathematics, digital and technical education, helping to address the shortage of science, technology, engineering and mathematics (STEM) skills<br>• Create a new National Retraining Scheme that supports people to re-skill, beginning with a £64 million investment for digital and construction training" |
| France | "Human Capital<br><br>To ensure a smooth transition towards an AI-oriented economy, a thorough transformation of learning paths is needed, involving both reforms to the initial education of upcoming generations and opportunities of vocational training and lifelong learning for the current and upcoming workforce.<br><br>The AI for Humanity strategy highlights two important prerequisites for the successful development of human capital in AI. A first prerequisite relates to the inclusion of effective and compulsory digital and AI-related disciplines at all levels of the education and training curricula. This requires both reforms to the course content and to the teaching methods used. A second prerequisite is that the proposed education pathways should be free of any social inequality. This could be achieved by setting up incentive policies to ensure more diversity and to achieve more equality in participation rates, with a special attention to counteract any form of gender stereotyping (e.g. by incentivizing participation of women into digital and AI courses)." |

**(Cont.) Table 5: Examples of explicit wording regarding education and social protection (if any) in national strategies**

| | |
|---|---|
| **Sweden** | "Training: AI creates an increased need for life-far learning. It is therefore necessary with opportunities for relevant continuing education and further education by already professionals." |
| **Germany** | "World of work and labor market: shaping structural change:<br><br>The potential for AI to serve society as a whole lies in its promise of productivity gains going hand in hand with improvements for the workforce, delegating monotonous or dangerous tasks to machines so that human beings can focus on using their creativity to resolve problems. This requires a proactive approach to the design of future of work"; "The draft legislation wants to give employees whose jobs are at risk of becoming lost to technologies, those otherwise affected by structural changes, and those wishing to train for a profession for which is labor is scarce, an opportunity to acquire the skills they need." |
| **Russia** | "The protection of human rights and liberties:<br><br>...ensuring the protection of the human rights and liberties guaranteed by Russian and international laws, including the right to work, and affording individuals the opportunity to obtain the knowledge and acquire the skills needed in order to successfully adapt to the conditions of a digital economy." |
| **Singapore** | "Adopt a human-centric approach<br><br>We will build an AI-ready population and workforce. At the societal level, as part of the overall promotion of digital literacy, we will raise awareness of AI, so that citizens are prepared for technological change, and are engaged in thinking about AI's benefits and implications for the nation's future. At the workforce level, we will prepare our professionals to adapt to new ways of working, in which workers are augmented by AI capabilities." |

**(Cont.) Table 5: Examples of explicit wording regarding education and social protection (if any) in national strategies**

For the US, the main preparation for the job disruption on the labor market is to: "Foster and educate active workers capable of leading the intelligent information society based on their creativity and emotional intelligence" to "ensure opportunities for a decent and humane standard of living by supporting the re-training of personnel and improving the employment and welfare environments." Similar content can be found in the AI strategy of Germany: "to give employees whose jobs are at risk of becoming lost to technologies, those otherwise affected by structural changes, and those wishing to train for a profession for which labor is scarce, an opportunity to acquire the skills they need." In the AI strategy of France, instead of guaranteeing a good level of living under AI, what is promised is only indiscriminating and equal access to re-training and education opportunities: "A second prerequisite is that the proposed education pathways should be free of any social inequality. This could be achieved by setting up incentive policies to ensure more diversity and to achieve more equality in participation rates, with a special attention to counteract any form of gender stereotyping (e.g., by incentivizing participation of women into digital and AI courses)."

We have also examined the national AI strategies of Western welfare states, such as Finland and Sweden, as well as Asian countries with Confucian tradition and family values such as China and Japan. These countries have been compared to others with regards to equity and social protection under AI job disruption. Surprisingly, little difference was found, meaning that a market-based and non-social-policy approach is the dominant and cross-cutting theme of most national AI strategies. For Finland, it states in its national strategy that: "The prerequisite for the broad-based utilization of artificial intelligence is that the population for the most part has a command of the skills and knowledge needed for its application. The requirements for the age of artificial intelligence should be visible in study content throughout the entire education system." Perhaps, what is even more surprising is, instead of assuring the provision of social protection in an AI society, it has pointed out the drawbacks and limitations of such schemes: "On the other hand, comprehensive earnings security insurance inevitably involves incentive problems. The long-term objective should be to increase the inventiveness of both social and unemployment security and improve the strengths related to these."

In China (an Asian, Confucian, and Socialist country), employment and re-training is still preferred to social protection: "Vigorously strengthen training for the labor force working in AI. Accelerate the study of how AI affects the employment structure, the change of employment methods and the skills demands of new occupations and jobs. Establish lifelong learning and employment training systems to meet the needs of intelligent economy and intelligent society, and support institutions of higher learning, vocational schools and socialization training institutions to carry out AI skills training, substantially increasing the professional skills of workers to meet the demands of the high-quality jobs in China's AI research."

Asian and Western countries typically have two different welfare state models in which the state in the latter provides a much better and a more generous protection of income and welfare to citizens (Aspalter, 2006). As a result, with Asian countries taking the same approach as the West in favoring education and re-training over improving social protection, the net impact of AI job disruption on the labor force could be much more extensive in Asia, with workers absorbing a higher share of the negative economic effects.

Despite the fact that the AI strategy of international organizations tend to be more prescriptive and guiding in nature, no significant differences between national AI strategies and those of international organizations regarding equity and social protection were found in our analysis. This means that taking a market-oriented approach and deploying re-training and education programs as a replacement for strengthening social protection is currently a well-accepted international norm. In the "OECD Principles on AI", it recommends: "Empower people with the skills for AI and support workers for a fair transition." In the EU White Paper on AI, it also recognizes "skills" as the most important hurdle in the transition to the AI society: "The European approach to AI will need to be underpinned by a strong focus on skills to fill competence shortages"; "Initiatives could also include the support of sectoral regulators to enhance their AI skills in order to effectively and efficiently implement relevant rules." In addition, "The Plan will also increase awareness of AI at all levels of education in order to prepare citizens for informed decisions that will be increasingly affected by AI."

Presumably, the use of market and re-training in lieu of an explicit social policy for addressing the job disruption builds on two tenets. First, it assumes that the market is self-regulating, and therefore could fix itself and take care of most issues and concerns including unemployment caused by AI job disruption. For example, the labor force could seek re-training opportunities by themselves or those opportunities would be provided by firms and employers. Second, a two-stage development strategy may be used in AI strategy. In the first stage, technological advancement and economic growth should be the main concern and focus. As society grows richer and accumulates more wealth through AI development, the government would have more resources to address the equity and social protection issues at a later stage. Nevertheless,

by past experience of technological change and international development, these two scenarios are more likely to be flawed and over-optimistic. For instance, labor in the "Danger Zone" might have limited access to re-training opportunities, and companies are likely to shift their investment to AI-rich countries rather than paying to train the labor force of a particular country. Training and education should also be public goods, which are mostly provided by the government, not by the market.

Since equity is one of the major market failures, relying on market self-adjustment alone for resolving equity issues is unrealistic and defies economic theory (Stiglitz, 2000). A country with plentiful resources being re-allocated through governmental actions (i.e., taxation and public expenditure) is not necessarily correlated empirically (Acemoglu & Robinson, 2012). Many studies have provided abundant evidence that economic inequalities persist in many well-developed and advanced countries (Aspalter, 2006). State-driven debates concerning the rise of AI should be complemented by a call for reform and modernization of the governmental apparatus and services to respond to the new needs of the digital society (Cheung, 2005; Dunleavy et. al., 2008). This leads to the conclusion that we should not assume the economic power of a nation would automatically translate into a fair and equitable society in the AI era.

## Conclusion: The Future Direction - Policy Gap and Recommendations

As we approach the era of AI job disruption, there is a policy gap between the demand for policy solutions and the supply of the current wealth of knowledge on the future of work. While there is a large amount of research and discussion on the impact of AI on economic growth and employment, there is relatively less research on what governments should do to turn the risk and threat of AI into job opportunities and social good for all. On the principle of "rise with AI, not race with it" (World Bank, 2018), governments must play an active or even aggressive role not only on economic growth and national competitiveness, but also on social protection and a fair re-allocation

of resources. However, this paper finds that many countries, especially developing ones, are not well-prepared for AI, and most countries seem to be overlooking fairness and equity issues under job disruption. The ideal state of AI will not be realized without a certain amount of effort, and the absence of proper policies and enabling factors could easily lead to a "AI Divide" between AI-rich countries and AI-poor countries. Policymakers must work hard to ensure those enabling factors, which include institutions and societal conditions, do exist for making sure their governments and countries are well prepared for the arrival of AI and its major impact on society, turning all possible threats into opportunities in order to bring progress and prosperity.

As revealed by analyzing various national AI strategies, focusing only on economic growth and national competitiveness whilst ignoring equity and social protection is a flawed and dangerous proposition. The proposition has an implicit assumption that as long as more wealth can be created by AI, equity issues can be resolved easily and over a certain period of time. The implicit assumption has overlooked a few very major and important points. First, equity is a market failure, and inequalities exist even in rich societies so that the role of the government in ensuring equity and fairness under AI job disruption is necessary. Second, as education and re-training have positive externalities and can even be taken as a "public good", a major and targeted investment headed by the government on education and re-training is necessary. In addition, some segments of the population may be vulnerable and cannot be easily retrained for AI (e.g., the older population). To them, new social protection programs such as UBI may be the best and only solution.

It is noteworthy that UBI was first raised in the book "Utopia" (1516) by Sir Thomas More, who also coined the word. In this connection, in the era of AI and job disruption, policies of equity and social protection would determine the difference between a utopian and dystopian future. It can further draw the line between job destruction or creative destruction. "Creative destruction" is the concept proposed by the famous economist Joseph Schumpeter (1942) which refers

to the process of industrial mutation that incessantly revolutionizes the economic structure from within to create a new and better one with more opportunities and resources. Paradoxically, without a reinforced state's role on equity and social protection, we can only see the destruction of jobs but never the creation of new opportunities brought by innovation and technology.

In the hybrid governance analyzed in this paper, it is hard to disentangle efforts to steer national policies in a particular direction from business interests. It seems rather unfortunate to see that job disruption and counteracting policies—especially towards those who may not be able to adjust—are all missing in the AI strategies of major countries. The change of new technology in AI requires the changing role of the state—including new capacities and integrated functions, allowing for fairness and equity. Future studies expanding on the knowledge frontier of the societal impacts of AI automation should pave the way towards understanding the intended and unintended consequences of the disruptive changes and shift of power brought by AI. In this regard, three major policy recommendations are made in the following.

## Recommendation 1: Theory and Practice

Governments should have more alignment and integration between theory and policy in formatting their AI strategies. Only by breaking the wall between academic research and policy discussion can there be a possibility for the formulation of effective policies well-supported by research and well-grounded in knowledge and theories. For example, governments should discuss how to prepare their labor force to rise with AI by equipping them with skills and capacities to work with enabling technologies rather than replacing technologies. Education and training in schools and the labor force should put more emphasis on social intelligence and creative intelligence, which are not going to be replaced by AI in the future of work.

## Recommendation 2: International Organizations and Developing World

AI impacts both developed and developing countries. That said, many developing countries are ill-prepared due to limitations in resources, technology know-how and policy capacity. National AI strategies have only been released by developed countries and global powers; no developing countries have set up a comprehensive AI strategy. Context and institutions also matter in determining the ability of a nation to embrace and survive job disruption by AI. Unlike the welfare states of Western countries, the social protection system of many developing countries is feeble and depends much more on self-reliance, the vitality of the economic system, and family support. This means that the ability of individuals to sustain economic instability and downturn caused by AI job disruption would be weak and non-sustainable. Understanding the limited capacities and resource concerns of developing countries, it is recommended that global and international organizations such as the World Bank, UN, and World Economic Forum take the lead in offering advice and support for developing countries to craft their own AI strategies.

## Recommendation 3: AI for All

A good AI policy should ensure that all members of society benefit from this powerful technology. To build on the major theme of "AI for Social Good", there should also be "AI for All" − benefiting and empowering all members in society. It is inevitable that some people, especially the older population, will likely find it difficult to re-train for the AI era. As society gets richer and wealthier with AI, how this vulnerable population should be protected and funded will require some tough decisions, which can be delayed but never avoided. In this connection, equity, social security, and fair re-distribution (e.g., introducing UBI to protect the vulnerable population) should be critical and essential elements in all future AI policy responses.

# References

Acemoglu, D., & Autor, D. (2011). Skills, Tasks and Technologies: Implications for Employment and Earnings. *Handbook of Labor Economics*.

Acemoglu, D., & Robinson, J. A. (2012). *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. NY: Crown Business.

Grönlund, Å. (2011). Connecting E-government to Real Government – The Failure of UN E-participation Index. *International Conference on Electronic Government*, 26-37.

Anderson, J. (2015). *Public Policy-Making*. Stamford: Cengage.

Asia Development Bank. (2018). *Asian Development Outlook 2018: How Technology Affects Jobs*. Asia Development Bank.

Aspalter, C. (2006). The East Asian Welfare Model. *International Journal of Social Welfare*, 290-301.

Bower, J. L., & Christensen, C. M. (1995). Disruptive Technologies: Catching the Wave. *Harvard Business Review, 73*(1), 43-53.

Cairney, P. (2016). *The Politics of Evidence-Based Policy Making*. NY: Palgrave Macmillan.

Cath, C. (2018). Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges. *Philosophical Transactions of the Royal Society A*.

Cheung, A. (2005). The Politics of Administrative Reforms in Asia: Paradigms and Legacies, Paths and Diversities. *Governance, 18*(2), 257-282.

Deming, D. J. (2017). The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics, 132*(4), 1593-1640.

Desouza, K. C. (2018). Delivering Artificial Intelligence in Government: Challenges and Opportunities. Washington. *IBM Center for The Business of Government*.

Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2008). *Digital Era Governance: IT Corporations, the State and E-government*. NY: Oxford University Press.

Evans, P. (1995). *Embedded Autonomy: States and Industrial Transformation*. NJ: Princeton University Press.

Ferro, E., N. Loukis, E., Charalabidis, Y., & Osella, M. (2013). Policy Making 2.0: From Theory to Practice. *Government Information Quarterly*, 359-368.

Florida, R. (2002). *The Rise of the Creative Class*. NY: Basic Books.

Fountain, J. (2001). *Building the Virtual State: Information Technology and Institutional Change*. Washington D.C.: Brookings Institution Press.

Frey, C. B., & Osborne, M. A. (2017). The Future of Employment: How Susceptible are Jobs to Computerization? *Technological Forecasting and Social Change, 114*, 254-280.

Haggard, S. (1990). Pathways from the Periphery: *The Politics of Growth in the Newly Industrializing Countries*. NY: Cornell University Press.

Haque, M. S. (1996). The Contextless Nature of Public Administration in Third World Countries. *International Review of Administrative Sciences, 62*(3), 315-329.

Howlett, M., & Ramesh, M. (1998). Policy Subsystem Configurations and Policy Change: Operationalizing the Postpositivist Analysis of the Politics of the Policy Process. *Policy Studies Journal, 26*(3), 466-481.

Johnson, C. (1982). *MITI and the Japanese Miracle: The Growth of Industrial Policy, 1925-1975*. CA: Stanford University Press.

Kang, J., & Francisco, J. P. (2019). Automation and the Future of Work in Developing Countries. *Transformation of Work in Asia-Pacific in the 21st Century*.

Kaplan, J. (2016). *Artificial Intelligence: What Everyone Needs to Know*. NY: Oxford University Press.

Kingdon, J. (1984). *Agendas, Alternatives, and Public Policies*. NY: Harper Collins.

Kitchin, R. (2014). The Real-Time City? Big Data and Smart Urbanism. *GeoJournal*, 1-14.

Krueger, A. O. (1974). The Political Economy of the Rent-Seeking Society. *The American Economic Review, 64*(3), 291-303.

Lee, J., & Moon, M. J. (2019). Coming Age of Digital Automation: Backgrounds and Prospects. *Transformation of Work in Asia-Pacific in the 21st Century*.

Lee, K.-F. (2018). *AI Super-Powers: China, Silicon Valley and the New World Order*. NY: Mariner.

Lindblom, C. (2004) "The Science of Muddling Through" in Jay Shafritz, Albert Hyde, and Sandra Parkes, eds., *Classics of Public Administration*. 5th ed. Belmont, CA: Wadsworth.

Norris, P. (2012). *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*. NY: Cambridge University Press.

North, D. (1990). *Institutions, Institutional Change and Economic Performance*. NY: Cambridge University Press.

OECD. (2018). Putting Faces to the Jobs at Risk of Automation. *Policy Brief on the Future of Work*.

OECD. (2019a). Artificial Intelligence in Society. *OECD*.

OECD. (2019b). Employment Outlook 2019: The Future of Work. *OECD*.

Olson, M. (1982). *The Rise and Decline of Nations*. New Haven: Yale University Press.

Partnership for Public Service. (2018). Using Artificial Intelligence to Transform Government. *IBM Center for the Business of Government*.

Partnership for Public Service. (2019). More Than Meets AI. *IBM Center for the Business of Government*.

Painter, M., & Pierre, J. (2005). Unpacking Policy Capacity: Issue and Themes. In *In Challenges to State Policy Capacity: Global Trends and Comparative Perspectives* (pp. 1-18). UK: Palgrave Macmillan.

Polidano, C. (2001). Don't Discard State Autonomy: Revisiting the East Asian Experience of Development. *Political Studies, 49*, 513-527.

Pollitt, C., & Bouckaert, G. (2011). *Public Management Reform: A Comparative Analysis – A New Public Management, Governance, and the Neo-Weberian State*. UK: Oxford University Press.

Przeworski, A., & Limongi, F. (1993). Political Regimes and Economic Growth. *Journal of Economic Perspectives, 7*(3), 51-69.

Radu, R. (2019). *Negotiating Internet Governance*. Oxford: Oxford University Press.

Rodrik, D. (1992). Political Economy and Development Policy. *European Economic Review, 36*(2-3), 329-336.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

Schumpeter, J. (1942). *Capitalism, Socialism and Democracy*. NY: Harper & Brothers.

Stiglitz, J. (2000). *Economics of the Public Sector*. NY: W. W. Norton.

Straub, E. T. (2009). Understanding Technology Adoption: Theory and Future Directions for Informal Learning. *Review of Educational Research*, 625-649.

Tam, K. Y. (2019). Digital Transformation in the 21st Century: Implications and Policy. *Transformation of Work in Asia-Pacific in the 21st Century*.

Welch, E. W., Hinnant, C. C., & Moon, M. J. (2005). Linking Citizen Satisfaction with E-Government with Trust in Government. *Journal of Public Administration Research and Theory, 15*(3), 371-391.

Welch, E., & Wong, W. (1998). Public Administration in a Global Context: Bridging the Gaps of Theory and Practice between Western and Non-Western Nations. *Public Administration Review, 58*(1), 40-49.

The World Bank. (2018). The Future of Work: Race with – not against – the Machine. *Research and Policy Briefs, World Bank Malaysia Hub No. 16*.

World Economic Forum. (2018). The Future of Jobs Report 2018. *World Economic Forum*.

World Intellectual Property Organisation. (2018). China Drives International Patent Applications to Record Heights; Demand Rising for Trademark and Industrial Design Protection. *WIPO.*

Wong, W. (2013). The Search for a Model of Public Administration Reform in Hong Kong: Weberian Bureaucracy, New Pubic Management or Something Else. *Public Administration and Development, 33*(4), 297-310.

Wong, W., & Welch, E. (2004). Does E-Government Promote Accountability? A Comparative Analysis of Website Openness and Government Accountability in Fourteen Countries. *Governance*, 275-297.

Wong, W., Welch, E., & Moon, M. (2006). What Drives Global E-Governance: An Exploratory Study at a Macro Level. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 275-294.

Yahya, F. B. (2019). Preparing the Future Workforce – Reskilling, Retraining and Redeploying and the Transformation of the Education System. *Transformation of Work in Asia-Pacific in the 21st Century*.

# Bios of Project Authors

### BASU, Arindrajit, Centre for Internet & Society, India

Arindrajit Basu is a research manager at the Centre for Internet & Society, India, where he focuses on the geopolitics and constitutionality of emerging technologies. He is a lawyer by training and holds a BA, LLB (Hons) degree from the National University of Juridical Sciences, Kolkata, and an LLM in public international law from the University of Cambridge, U.K.

### BENTLEY, Caitlin, Australian National University & Sheffield University

Caitlin Bentley joined the 3A Institute in 2018, and assisted in the development of a new branch of engineering through her role as Research Fellow teaching the 3Ai Master of Applied Cybernetics until 2020. She is now a lecturer in AI-enabled Information Systems at Sheffield University's iSchool. Caitlin conducts research on cyber-physical systems, and how to make them more socially inclusive. With a research career spanning from Canada to the UK, Africa, Southeast Asia and Australia, Caitlin has contributed to a number of projects focused on enhancing learning and accountability through ICT, open development, the platform economy and artificial intelligence. Caitlin holds a PhD in Human Geography from Royal Holloway University of London, UK, an MA in Educational Technology from Concordia University, Canada, and a BA in Computer Science from McGill University, Canada.

### FINDLAY, Mark, Singapore Management University

Mark Findlay is a Professor of Law at Singapore Management University, and Deputy Director of its Centre for AI and Data Governance. In addition, he has honorary Chairs at the Australian National University, and the University of New South Wales. Professor Findlay is the author of 29 monographs and collections and over 150 refereed articles and book chapters. He has held Chairs in Australia, Hong Kong, Singapore, England and Ireland. or over 20 years he was at the University of Sydney as the Chair in Criminal Justice, the Director of the Institute of Criminology. Most recent publications include Law's Regulatory Relevance and Principled International Criminal Justice: Lessons from tort law.

### HICKOK, Elonnai, Centre for Internet & Society, India

Elonnai Hickok is Chief Operating Officer at the Centre for Internet & Society India (CIS). Elonnai has graduated from the University of Toronto where she studied international development and political science. Elonnai leads the privacy, surveillance and cyber security work at the Centre and has also written extensively on issues pertaining to intermediary liability, digital rights, identity, cyber security and DNA profiling.

### HONGLADAROM, Soraj, Chulalongkorn University

Soraj Hongladarom is Professor of Philosophy and Director of the Center for Ethics of Science and Technology at Chulalongkorn University in Bangkok, Thailand. He has published books and articles on such diverse issues as bioethics, computer ethics, and the roles that science and technology play in the culture of developing countries. His concern is mainly on how science and technology can be integrated into the life-world of the people in the so-called Third World countries, and what kind of ethical considerations can be obtained from such relation. His most recent book is The Ethics of AI and Robotics: A Buddhist Viewpoint, forthcoming this year from Rowman & Littlefield. He is also the author of The Online Self and A Buddhist Theory of Privacy, both published by Springer. His articles have appeared in The Information Society, AI & Society, Philosophy in the Contemporary World, and Social Epistemology, among others.

### LEE, Kyoung Jun, Kyung Hee University

Kyoung Jun Lee is a professor of Kyung Hee University. He graduated with a BS/MS/PhD in Management Science from KAIST, and completed a MS/PhD course in Public Administration at Seoul National University. He won Innovative Applications of Artificial Intelligence Awards in 1995, 1997, and 2020. He was a visiting scholar at CMU, MIT, and UC Berkeley. He is 2017's president of Korean Intelligent Information Systems Society and current president of Korean Association for Business Communication. He is currently the Director of Big Data Research Center and International Center for Electronic Commerce. He received 2017 Presidential Award for e-government of Korea.

### MOON, M. Jae, Yonsei University

M. Jae Moon is Dean of the College of Social Sciences and Director of the Institute for Future Government at Yonsei University. His research interests include digital government, public management, and comparative public administration. He is an elected Fellow of National Academy of Public Administration (NAPA). Recently, he was recipient of the Highest Research Award of Yonsei University in 2019 and the Stone Award of the American Society for Public Administration in 2020. He is also selected as one of world's 100 most influential people in Digital Government 2018 and 2019 consecutively by Apolitical which is a London-based leading non-profit organization.
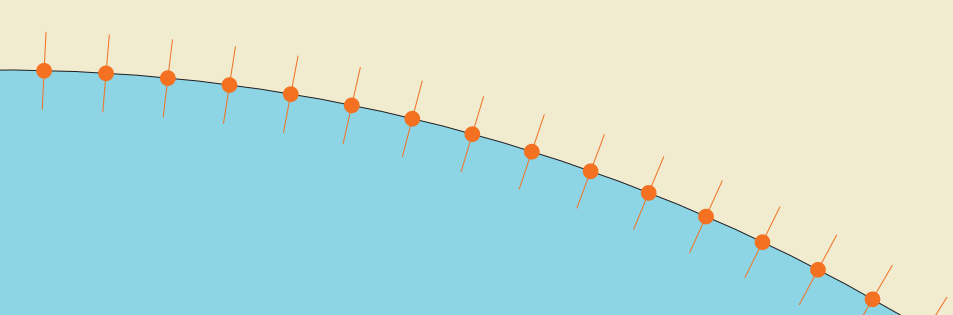
### SINHA, Amber, Centre for Internet & Society, India

Amber Sinha is the Executive Director at the CIS. He works on issues surrounding privacy, big data, and cyber security. Amber is interested in the impact of emerging technologies like artificial intelligence and learning algorithms on existing legal frameworks, and how they need to evolve in response. He has studied humanities and law at National Law School of India University, Bangalore.

### WONG, Wilson, Chinese University of Hong Kong

Wilson Wong is an Associate Professor of the Department of Government and Public Administration and the Director of the Data Science and Policy Studies (DSPS) Programme of the Social Science Faculty in the Chinese University of Hong Kong. He received his bachelor degree in the Chinese University of Hong Kong, a Master in Public Administration degree and a PhD in Public Administration, both from Syracuse University. Professor Wong's major areas of research include ICT and E-governance, Big Data, AI and public policy, public management and comparative public policy. He has served as a visiting fellow of the Brookings Institution and Harvard University.

### YARIME , Masaru, Hong Kong University of Science and Technology

Masaru Yarime is an Associate Professor at the Division of Public Policy in the Hong Kong University of Science and Technology. He also has appointments as Honorary Reader at the Department of Science, Technology, Engineering and Public Policy in University College London and Visiting Associate Professor at the Graduate School of Public Policy in the University of Tokyo. He received BEng in Chemical Engineering from the University of Tokyo, MS in Chemical Engineering from the California Institute of Technology, and PhD in Economics and Policy Studies of Innovation and Technological Change from Maastricht University in the Netherlands.

# Acknowledgement

# Partnership

We in particular thank United Nations ESCAP and Google for their continuing support to this project.
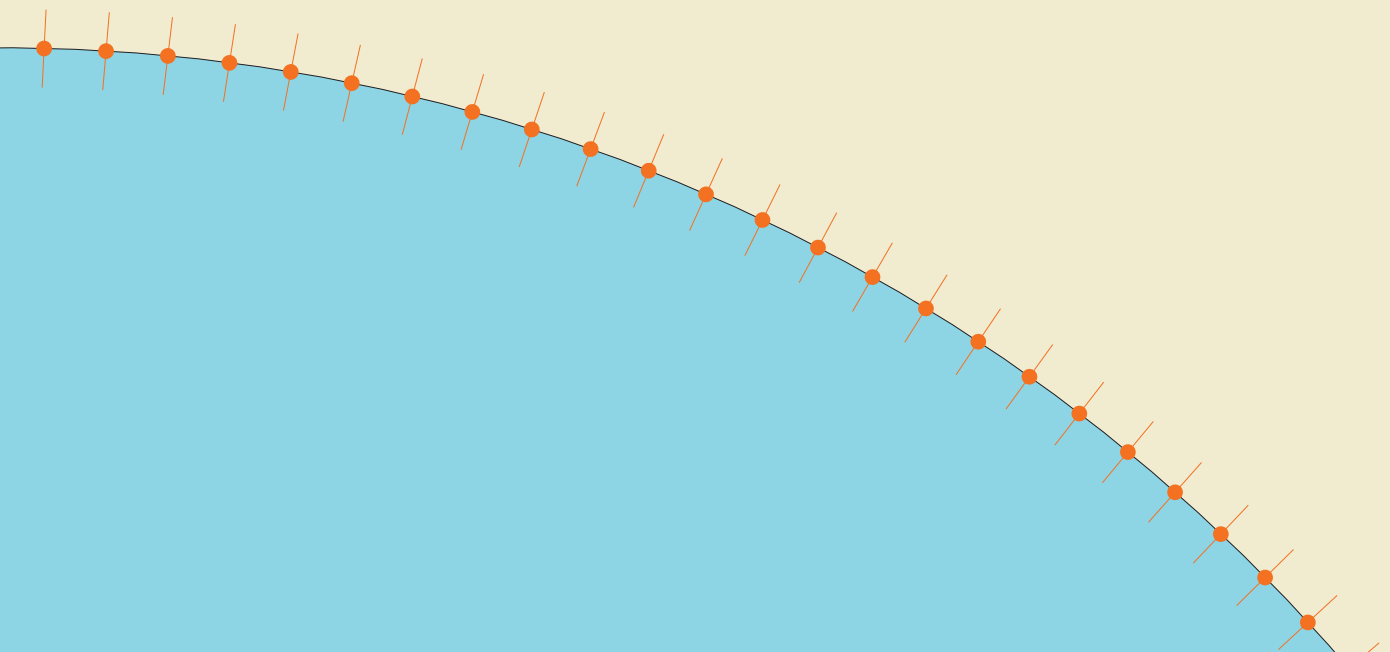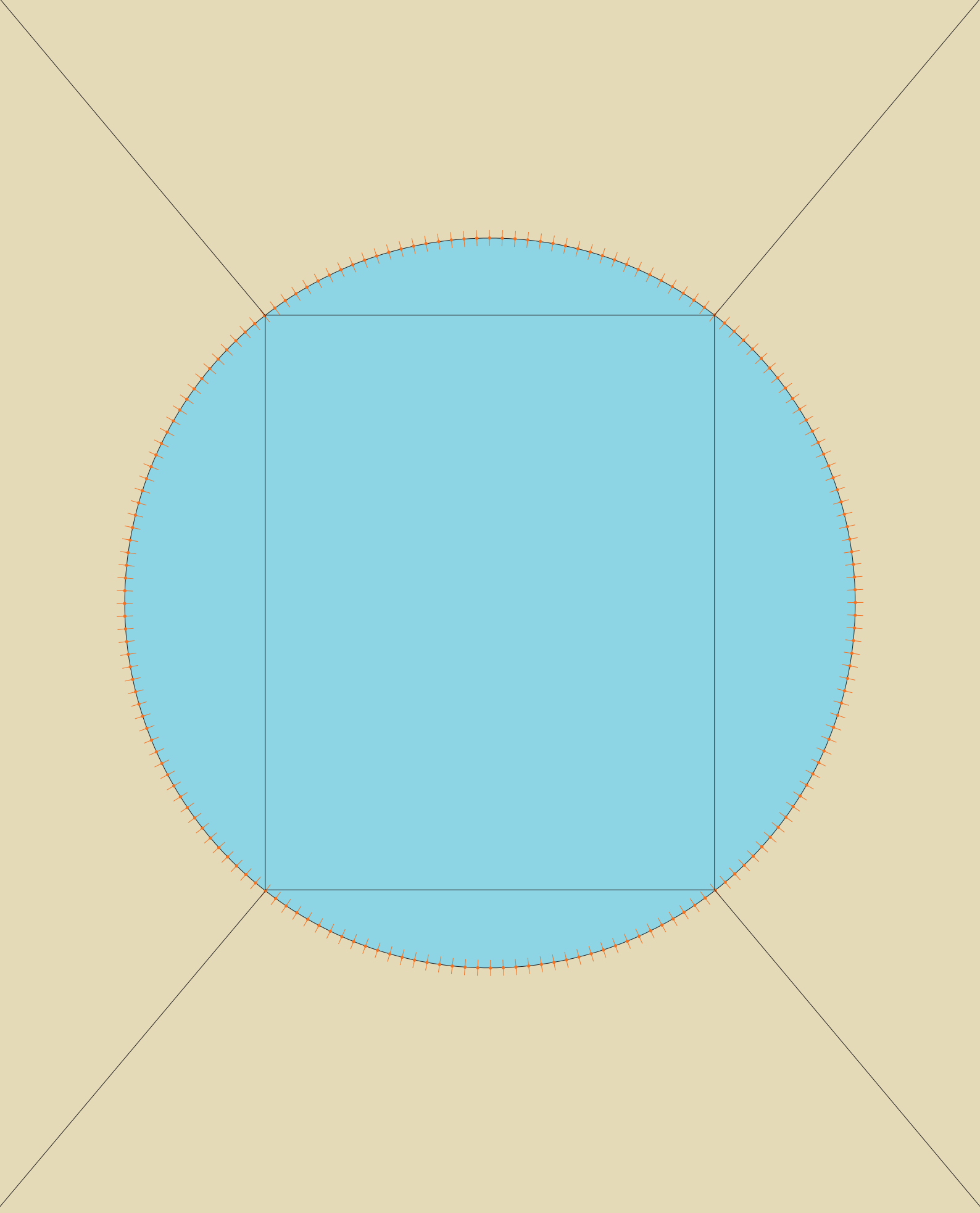
## United Nations ESCAP

The Economic and Social Commission for Asia and the Pacific (ESCAP) serves as the United Nations' regional hub promoting co-operation among countries to achieve inclusive and sustainable development. It is the largest regional intergovernmental platform with 53 Member States and 9 Associate Members. The Commission's strategic focus is to deliver on the 2030 Agenda for Sustainable Development, through reinforcing and deepening regional co-operation and integration to advance connectivity, financial co-operation and market integration. ESCAP, through its research and analysis, policy advisory services, capacity building and technical assistance, aims to support sustainable and inclusive development in member countries.

## Google

Google's mission is to organize the world's information and make it universally accessible and useful. We believe that AI is a powerful tool to explore and address difficult challenges such as better predicting natural disasters, or improving accuracy of medical diagnoses. In 2018, we launched AI for Social Good to meaningfully contribute to these solutions, drawing on the scale of our products and services, investment in AI research, and our commitment to empowering the social sector with AI resources and funding.