Digital Futures Lab | Konrad Adenauer Stiftung

Reframing AI Governance: Perspectives from Asia

URVASHI ANEJA

RAMATHI BANDARANAYAKE

JENNIFER BOURNE

JULIA CHEN (陈英)

YUCHEN CHEN

MARK FINDLAY

MAYA INDIRA GANESH

CINDY LIN

VIDUSHI MARDA

JUN-E TAN

WILLOW WONG

digital futures lab

KONRAD ADENAUER STIFTUNG

Digital Futures Lab |
Konrad-Adenauer-Stiftung

This book is supported by the R**ule of Law Programme, Konrad-Adenauer Stiftung.**

Aneja, U (Ed.). (2022). *Reframing AI Governance: Perspectives from Asia*. Digital Futures Lab; Konrad-Adenauer-Stiftung.

Publication Support:
Aishwarya Natarajan

Production Lead:
Sasha John

Design Lead:
LMNO Design / www.lmno.in

Table of Contents

Abbreviations

ARMP
Algorithmic recommendation management provisions

API
Application Programming Interface

AI
Artificial Intelligence

AICx
Artificial Intelligence Centre of Excellence

BDA
Big Data Analytics

BPPT
Badan Pengkajian Penerapan Teknologi/Agency for Application and Assessment of Technology

BAAI
Beijing Academy of Artificial Intelligence

CDR
Call Detail Records

CIS
Case Information System

CCCDR
Central Committee for Comprehensively Deepening Reform

CAICT
China Academy for Information and Communications Technology

CCP
Chinese Communist Party

CCTNS
Crime and Criminal Tracking Network System

CAC
Cyberspace Administration of China

FRT
Facial Recognition Technology

4IR
Fourth Industrial Revolution

IBM Garage
IBM Garage Methodology

ICTA
ICT Agency of Sri Lanka

IP
Intellectual Property

IoT
Internet of Things

ICJS
Interoperable Criminal Justice System

MyCC
Malaysia Competition Commission

MDEB
Malaysia Digital Economy Blueprint

MDEC
Malaysia Digital Economy Corporation

KKMM
Ministry of Communications and Multimedia

MITI
Ministry of International Trade and Industry

MOST
Ministry of Science and Technology

MOSTI
Ministry of Science, Technology, and Innovation

MSC
Multimedia Super Corridor

N4IRP
National Fourth Industrial Revolution Policy

AI-RMap
National AI Roadmap

AFRS
National Automated Face Recognition System

NDSO
National Database on Sexual Offenders

NDHGS
National Digital Health Guidelines and Standards

NFCP
National Fiberisation and Connectivity Plan

Industry4WRD
National Policy on Industry 4.0

PIKOM
National Tech Association of Malaysia

AIDP
New Generation AI Development Plan

OGP
Open Government Partnership

PIPL
Personal Information Protection Law

RTI
Right To Information

S&T
Science and Technology

SF
Science Fiction

SCS
Social Credit System

SLASSCOM
Sri Lanka Association for Software and Services Companies

CCPPC
Subcommittee of Economy of National Committee of the Chinese People's Political Consultative

SUVAS
Supreme Court Vidhik Anuvaad Software

TISL
Transparency International Sri Lanka

UGC
University Grants Commission

UTM
Universiti Teknologi Malaysia

Foreword

STEFAN SAMSE
Director, Rule of Law Programme Asia

Surveys from leading business consulting firms repeatedly indicate that Asia is all set to emerge as a powerhouse for Artificial Intelligence (AI) technology. Over the last five years, we have seen a sharp rise in the number of Asian countries coming out with their national AI strategy documents. South and Southeast Asia in particular have already emerged as strong bases for the technology start-up ecosystems. The governments in this region have been actively setting up AI strategies to reap the full benefits of technology to ensure greater economic development for their societies.

In theory, the aim of these governments are praiseworthy but the path to development and deployment of AI technology is not without its hurdles. The extensive and unregulated development and deployment of technology has raised several issues in the region including but not limited to use of surveillance technologies, algorithms deepening societal faultlines, increased and unfettered power of technology companies in these markets. While AI is projected as a panacea for the economic difficulties that plague the region, the governments have to think deeply about the limitations of these technologies. Importantly, the role of the state in regulating AI technology requires more critical attention. China and the European Union have already introduced their AI governance regime in the last two years. The current understanding of the concepts of algorithmic fairness, accountability, transparency, and ethics have all emerged from the Western social context. In order to effectively establish a regulatory framework for AI governance, Asian States have to think more deeply about what these concepts mean in their local context and what are the ways in which existing social thought can be converted into policy.

It is in this context that we bring forth this publication, which captures the perspectives from South and Southeast Asia about AI governance in the region. In 2018, the worldwide Rule of Law Programme of the Konrad-Adenauer-Stiftung (KAS) spearheaded an initiative to look at the impact of digitalisation on our society; more specifically, the initiative was desirous of exploring the linkages between law, policy and technology. In line with this goal, the Rule of Law Programme Asia took infant steps to set up the TechLaw programme for the Asian region. We truly hope that this publication is a valuable contribution in making voices from Asia heard in the international policymaking arena.

Socio-Material Steering of AI Innovation and Governance

Konrad-Adenauer-Stiftung

Digital Futures Lab |

Socio-Material Steering of AI Innovation and Governance

URVASHI ANEJA

Machine learning and AI tools increasingly permeate vast areas of political, economic, and social life. The use of AI systems can enable efficiency and productivity gains but is also behest by a range of complex problems and challenges. Developing strategies to align AI development and deployment trajectories with social justice is thus an urgent policy priority.

Most frameworks for AI governance tend to coalesce around a common set of principles, such as transparency, accountability, privacy, equality, and safety. While these principled frameworks provide a normative mooring to AI innovation trajectories, they can also be interpreted in multiple ways and are hard to operationalise. For the most part, ethical frameworks seem like a way for industry to argue that self-regulation will be enough to check AI harm and avoid other command and control forms of regulation. Concerned with regulating too early or regulating too much, many governments are also developing such principled frameworks to support self-regulation by technology companies themselves. Even in the case of the European Union, whose policies around AI are some of the most mature, only high-risk cases require stringent regulatory interventions. However, risk-based assessments are necessarily reductive and subjective and even low-risk risk cases can undermine human rights and have harmful structural impacts.[1]

A more promising direction is the turn toward the application of international human rights frameworks. They are universal and binding and codified in international law; the responsibilities of governments and companies are well-articulated, and a

1 European Center for Not-For-Profit Law, "*Evaluation the Risk of AI Systems to Human Rights from a Tier-Based Approach*", ecnl.org, March 23, 2021. https://ecnl.org/news/evaluating-risk-ai-systems-human-rights-tier-based-approach.

range of domestic, regional, and international institutions are available to provide remedy. But rights-based frameworks take the individual as the primary unit of concern, whereas harm, in an AI-driven world, is often collective and structural. A single instance of harm could be unrecognisable and inconsequential, but many taken together could have harmful consequences in the aggregate, and over a period alter societal structures and relationships.[2]

Common to these emerging approaches to AI governance is that they tend to be dominated by the experiences and priorities of a select few industrialised economies. However, the challenges, opportunities and harms posed by AI systems are likely to differ across social, economic, and cultural contexts. Most countries in the global south are still struggling to bridge the digital divide, even while digital technologies are deemed pivotal to addressing long-standing development challenges. Weak state and institutional capacities further hinder the ability of governments to effectively govern emerging technologies like AI. It also cannot be assumed that terms like fairness, transparency or accountability carry the same meaning across socio-political contexts in the global south. Countries are also at different stages of socio-economic development and occupy different places in the global AI value chain. Understanding AI as a socio-technical system thus requires us to think beyond managing its technological affordances, and contend with how power, interests, and values across different socio-cultural contexts shape AI trajectories and are reconfigured in the process.

This volume seeks to understand perspectives on AI governance from South and South-East Asia. What are the politics, values, institutions, and policies shaping AI governance across Asian countries? How well do global conversations around AI governance translate to Asian contexts, with differing state priorities, institutional capacities, and cultural contexts? Are there alternative frames or discourses around AI governance emerging from the Asian region?

Governance options are deeply connected to the way a problem gets framed – it shapes how issues are prioritized and negotiated, rendering some policy options desirable while restricting others. This volume of essays highlights the various frames through which governments approach AI governance in select Asian countries and its implications for policy prioritization and intervention. The chapters also show how local civil society actors are negotiating dominant 'western' frames to reflect their social priorities and cultural histories. The volume further emphasizes the necessity of developing new or alternative frames for AI governance so that AI innovation can

2 Anna Lauren Hoffman, "*Where fairness fails: data, algorithms and the limits of antidiscrimination discourse*", Taylor & Francis Online, May 13, 2019. https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1573912?journalCode=rics20

enhance, rather than undermine, bonds of community and social trust and prioritize the experiences of those most impacted by the use of AI systems

> “Understanding AI as a socio-technical system thus requires us to think beyond managing its technological affordances, and contend with how power, interests, and values across different socio-cultural contexts shape AI trajectories and are reconfigured in the process.”

Chapters in this Volume

The first three chapters in this book show how governments in Sri Lanka, Malaysia and China frame the issue of AI governance. While these countries are at different stages of AI policy, they similarly view AI through the prism of economic growth and development – the focus of policy is to create an enabling environment for AI development and deployment.

Ramathi Bandaranayake’s chapter on **‘The Beginnings of AI and Data Governance: The Landscape in Sri Lanka’** looks at the issue of AI governance from the perspective of a small resource-constrained country. It shows how AI deployment and governance are still at a nascent stage in Sri Lanka. Sri Lanka is yet to put forward a policy framework specific to AI and the limited conversation around AI is led by the private sector, industry bodies and academia. Much of the current policy focus is on enabling data access and data protection. Sri Lanka’s data protection law, the first country in South Asia to have one, does not address the issue of AI specifically but gives data subjects the right to review decisions based solely on automated processing. A National Data Sharing Policy has also been introduced, but the unavailability of high-quality, machine-readable datasets remains a major constraint to developing localised solutions, with researchers and developers tending to use data from other national contexts. Issues such as privacy have begun to receive more attention with the use of digital technologies to manage the COVID-19 pandemic, but broader harms and risks entailed in the use of AI such as discrimination, inequality, and misuse, are yet to enter policy conversations. This means that there is an opportunity to learn from other countries and build norms and frameworks early on to prevent privacy violations and discriminatory outcomes.

Jun-E Tan’s chapter, **‘To What Extent Does Malaysia’s National Fourth Industrial Policy Address AI Security Risks’,** unpacks the frames through which AI policy is approached in Malaysia. AI strategy is framed as part of the National Fourth

Industrial Revolution Policy (N4IRP); AI is considered the 'electricity' of the N4IRP and a key driver in 'transforming the socio-economic development of the country'. The priority is on economic growth and most of the action plans in the N4IRP focus on economic security, with a business-friendly, business-as-usual approach. As a result of this frame, political and social risks, such as misinformation or surveillance, are not considered in the policy. But because economic issues receive the most prominence in the policy, the N4IRP does include a discussion on issues such as labour displacement and social protection for gig workers. The policies are marked by a sense of inevitability and insecurity – inevitability about technological change and insecurity about being left behind. Together, these create a sense of urgency around the need to develop policies that allow Malaysia to leverage AI for economic growth, with a focus on speed instead of safeguards.

As Julia Chen's chapter on **'Chinese AI Governance: Past, Present and Future of Chinese AI Regulation'** shows, policy conversations on AI governance are the most advanced in China. Between 2017 and 2020, Chinese authorities adopted a 'wait and see' approach of trusting the market, encouraging innovation, and strengthening regulation when incidents happen. But as the inadequacy of self-regulatory measures became apparent, the government began to adopt stronger regulatory measures. Between 2020-21, multiple departments began taking stronger action against AI harms, such as the prohibition of facial recognition technologies without consent in commercial spaces, bans on differential pricing, and regulations on algorithmic recommendation systems (the first of its kind globally). This shift has been at least partly shaped by policymakers' desire to restrain the 'wild growth' of internet platforms. Keeping these platforms in check is deemed as necessary if China is to maintain political stability and the foundation of the Party's rule: realise common prosperity and reduce economic inequality. But despite such tighter regulation, the government remains committed to leveraging the economic and social potential of AI and is thus keen to promote its targeted deployment. This nuanced position is visible in data and AI legislation in Shenzen. Passed in June 2021, Shenzhen's data regulations include measures to promote the integration and sharing of public data, while also placing limits on data collection of data and the deployment of algorithms.

We also need to ask why and how certain frames become dominant, what they obscure, and the actors and governance frameworks they legitimate. Vidushi Marda's chapter **'An Ill-Advised Turn: AI Under India's E-Court Proposal'** draws attention to priorities, processes, and actors that lead to the dominance of certain frames. She shows how a recent proposal to introduce AI and other emerging technologies into the judicial system gives into tendencies of technological solutionism. The chapter draws attention to the role of private sector players in defining the problems that AI needs to solve and links it to the broader conversation about private capture of democratic and legislative processes. She points to the hastiness of the Indian government in adopting AI solutions but provides the cavaeat that this is because of the over-reliance on private actors to suggest solutions and market technologies to the government in the absence of any meaningful civil society involvement. India

has released a national strategy on AI, including a framework for Responsible AI. The framework emphasises values of fairness, accountability, and transparency and proposes a risk-based approach to AI governance. But as Vidushi's chapter shows, these remain mostly principled guidelines with little attention to their operationalisation. Basic checks and balances, for example, are missing from the proposed policy for introducing AI in India's judicial system. The chapter calls for Indian policymakers to pause and reflect on the inherent nature of AI systems, their limitations and appropriateness in supplanting various State functions.

Julia's chapter on AI governance in China is also noteworthy in this regard because it points to the role of netizens, journalists and academics in shifting the frames through which AI governance is approached, something not often heard in dominant narratives on China. Harder regulatory measures were introduced in China at least partly because of civil society pressure. Local governments, industry bodies, and academic institutes are also beginning to play a more active role in monitoring and documenting harms.

Cindy Lin and Yuchen Chen's chapter on the **'Myth of Data-Driven Authoritarianism in Asia'** discuss the frames through which civil society actors view and strategically use AI in China and Indonesia. They show that AI and data-driven technologies are not simply tools to enact authoritarian governance in Asia, as often-depicted in Western media, but are also techniques to intervene in oppressive social, political, and economic conditions and ideologies. Discussing the Chinese social credit system, they argue that it should not only be seen as a top-down system of control. Rather it was initiated by citizens themselves, as a way to manage the twin goals of 'modernisation and social justice', to enable 'a harmonious society' in the context of a growing market economy and concerns about its harmful societal impacts.

Similarly, in Indonesia, agile software development methods, typically associated with a means of improving labour productivity through neo-liberal techniques of self-improvement and surveillance in the West, acquired a very different meaning. Junior engineers in state bureaucracies perceived agile software methods as a means to transform existing bureaucracies into equitable, transparent, and entrepreneurial organisations.

AI ethics has been constructed under 'western-bound regulatory frameworks', through principles such as accountability and transparency. But, as this chapter argues, we need a more situated ethics rooted in historical and cultural contexts to understand local negotiations and contestations.

Mark Findlay and Willow Wong's chapter on **'Kampong Ethics'** similarly questions the applicability of universalising ethical frameworks, mostly imported from the West, in addressing the ethical priorities of diverse communities around the world. They argue that the current landscape of ethical guidelines and frameworks from western technology companies and governments has been largely self-serving. The critical ethical challenge is not abstract principles like privacy or

accountability but the instrumental and wealth-oriented objectives of current technology development processes.

Instead, they argue that the kampong (village) spirit of solidarity and community can provide a more solid foundation for steering AI toward social good. The kampong spirit is not one single thing, but that which can effectively stimulate and maintain reciprocal bonds and shared trust within socially located communities. The focus of AI design, development, and regulation, they argue, should be the enhancement of community, understood as social bonds of trust. AI creators and policymakers should design new technologies in service of enhancing communal relationships within specific communities. Without the commitment to locate AI in communities, the rollout of AI can undermine existing trust relations.

Maya Indira Ganesh and Jennifer Bourne's chapter, **'Between Threat and Tool: The Poetics and Politics of AI Metaphors and Narratives in China'** highlights how metaphors and narratives provide an entry point to understanding regional and cultural values, assumptions, and beliefs around AI. Chinese development and application of AI have generated antagonist concerns in much western policy and media circles. But, by looking at the frames through which AI is discussed within the country's own cultural materials, their chapter shows that there is a great deal of similarity in the metaphors used in China and other countries. Based on a reading of award-winning science fiction, short stories, and business media from China, they identify a metaphor that is like that found in dominant western dominant narratives – that of AI as both a potential threat and a tool. In these works of fiction, AI is positioned as threatening to humans but eventually human intelligence triumphs. These give us insight into the interior condition experienced by humans enmeshed in and with the future of automation.

These stories exist alongside metaphors of AI in business advertising as a tool that will work for us. But this narrative of AI as a future tool obscures the labor that is already entailed in the production of AI – the people who constitute AI's workforce, like the online workers who tag and label images for computer vision, test drivers of autonomous vehicles, platform-based workers, and factory workers. The chapter concludes with an important provocation, asking readers to consider what metaphors the gig workers, and other forms of invisible labor entailed in the production of AI, might deploy to express their desires and concerns.

Crafting an agenda for research and policy

The chapters in this volume show there is considerable convergence between how governments are approaching the governance of Artificial Intelligence. Among Asian countries, AI is seen as necessary for achieving socio-economic development and growth. The focus of governance initiatives is thus to create an enabling environment

for AI development and deployment. Even beyond Asia, there seems to be some convergence between how China and the EU are approaching AI governance. In both geographies, government authorities are trying to reign in powerful digital platforms through stringent regulatory measures while at the same time trying to develop mechanisms through which public data and data controlled by private companies can be leveraged to promote socially beneficial outcomes. There is perhaps greater space for global dialogue than is conventionally thought and countries would do well to look at recent policy interventions in China as they think about their governance strategies.

But, at the same time, we must not assume that the meaning and impact of AI are the same across the globe. Dominant frameworks of AI governance assume a certain universality in their approach. But, as these chapters show, we need to pay far closer attention to the local particularities of AI development and deployment. Concepts such as privacy or surveillance, for example, may not have the same importance or meaning across diverse socio-cultural contexts and local actors may strategically use AI tools to pursue different ends than is conventionally understood. We need to invest far greater resources in understanding these local interactions.

The irony is that not only are current AI innovation and governance paradigms dominated by the experiences and priorities of a select few industrialized economies but also the critiques of AI and proposals for alternate innovation paradigms are led by research and civil society organizations from the same set of industrialized economies. This must change if we are to create more inclusive and equitable AI futures. Governments and funding agencies must invest in research and civil society capacities in developing countries. Becoming global leaders in AI must not be only about developing new products or unicorn companies but requires a broader, whole of society, approach. Countries in Asia would do well to learn from the investments made by industrialized economies in research institutions and civil society – these investments have been crucial to the emergence of these countries as global AI leaders.

The chapters also provoke us to question whether safe, equitable and socially beneficial outcomes are possible with the current paradigm of AI innovation and deployment. The evidence on the harms of AI is mounting – from the manipulation of electoral outcomes to the degradation of labor rights to the increasing incidences of mental trauma among youth. Current AI innovation trajectories have also resulted in the unparalleled concentration of power in the hands of a select few technology companies while also increasing the surveillance capacities of national governments. Technical fixes such as privacy by design or anonymization are only partially effective, at best, and principles such as explainability are difficult to realize, if not impossible, in deep learning systems. One may even argue that the current innovation paradigm represents a failure of imagination and is predicated on exploitative practices. The very possibilities of machine learning have been enabled by what Birch et. al call data rentiership – i.e., the process by which actors extract revenue by exercising control over data. Innovators are innovating to make and capture

economic rents through their control over data as an asset, rather than creating value through production.[3]

Finally, rather than only focus on the future promise of AI technologies – which in itself is questionable because there is enough evidence to show the brittleness and fallibility of AI technologies[4] – we should focus on the already existing socio-material impacts. This should be the anchor for AI innovation and governance paradigms.

Cheap labour from the global south is what is enabling the production of AI systems – whether in terms of extraction of natural resources needed to build AI hardware and infrastructure or the often-invisible work of annotating and labelling data sets.[5] Workers are paid a few cents an hour, have to work long and unpredictable hours, face discrimination, and have no control over the terms and conditions of their labor. Similarly, gig work is growing and is even seen as necessary to counter slow job growth – this is particularly the case in developing countries like India with high levels of unemployment and a large youth population. But gig workers are subject to opaque algorithmic management and monitoring systems resulting in low wages and degradation of labour agency and rights.[6] Studies also highlight the environmental impacts of AI. For example, training a single AI model emits CO2 comparable to that of 5 cars over their lifetimes.[7] Making AI models more accurate consumes even more electricity. A recent study quoted in HBR shows that the last 0.08% incremental increase in accuracy of an AI model took nearly 400% more energy than the first stage.[8]

We need to centre these socio-material conditions of AI innovation in conversations around AI governance. Centring these socio-material concerns will put necessary brakes on the pace of AI deployment, creating greater space for developing alternative innovation paradigms and strengthening governance and societal capacities.

3 Kean Birch, Margaret Chiapetta and Anna Artyushina, "*The problem of innovation in technoscientific capitalism: Data rentiership and the policy implications of turning personal digital data into a private asset*", Policy Studies, March 31, 2020. https://www.researchgate.net/publication/338489125_The_problem_of_innovation_in_technoscientific_capitalism_Data_rentiership_and_the_po icy_implications_of_turning_personal_digital_data_into_a_private_asset

4 Inioluwa Deborah Raji*, I. Elizabeth Kumar*, Aaron Horowitz, and Andrew D. Selbst, "*The Fallacy of AI Functionality*", 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022. https://arxiv.org/pdf/2206.09511.pdf

5 Kate Crawford, "*Atlas of AI*", Yale University Press, April 6, 2021. https://yalebooks.yale.edu/book/9780300264630/atlas-of-ai/

6 Worker Info Exchange, "*Managed by Bots: Data-Driven Exploitation in the Gig Economy*", workerinfoexchange.org, December, 2021. https://www.workerinfoexchange.org/wie-report-managed-by-bots

7 Karen Hao, "*Training a single AI model can emit as much carbon as five cars in their lifetime*", MIT Technology Review, June 6, 2019. https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

8 Sanjay Podder, Adam Burden, Shalabh Kumar Singh and Regina Maruca, "*How Green is Your Software?*", Harvard Business Review, September 18, 2020. https://hbr.org/2020/09/how-green-is-your-software

Because many of these technological transformations and policy frameworks are still at a nascent stage in Asia, there is an opportunity to use learnings from other countries to craft alternative AI innovation trajectories and avoid harmful technological and policy lock-ins. AI governance conversations are often limited to the high corridors of power. But, in many ways, this is a conversation about our shared and contested futures - what are our societal priorities or visions of a 'good life' and what role do we want automated systems to play in that life?

The COVID-19 pandemic has shown us that social production must precede economic production, that the latter is not possible without the former. Future innovation trajectories must build on this observation i.e. be directed towards those goals that nurture social production. The logic guiding current technological development and business models is contrary to the goals of social production – from rising incidences of mental health and online abuse to the fragmentation of workers through gig work platforms, to the disintermediation of government officials with the use of automated systems.

Instead, we could look to social justice movements around the world to understand peoples' needs, desires, and priorities. Such movements often bring together or speak of the concerns of the most marginalised members of our society. Our collective visions of the future must centre these concerns, or we will continue to reproduce existing patterns of societal inequity and injustice.

> "AI governance conversations are often limited to the high corridors of power. But, in many ways, this is a conversation about our shared and contested futures - what are our societal priorities or visions of a 'good life' and what role do we want automated systems to play in that life?"

CHAPTER 01

The Beginnings of AI and Data Governance: The Landscape in Sri Lanka

01

CHAPTER 01

The Beginnings of AI and Data Governance: The Landscape in Sri Lanka

RAMATHI BANDARANAYAKE

Abstract

While many governments around the world have recognized the potential of artificial intelligence (AI) and are developing policies and strategies to harness these benefits, many countries are still in the early stages of this process. This dynamic can be observed quite clearly in the South Asian region. The government of India has been the most visibly proactive in discussing AI and data governance and releasing numerous policy documents. Other countries, however, lag relatively behind. This chapter considers the experiences of Sri Lanka, ranked 90 in the 2020 Government AI Readiness Index, where the discourse on AI and data governance is still at a very embryonic stage. The chapter first discusses the current landscape of AI and data, some of the policy initiatives that have taken place, and finally, makes recommendations for how norms may be set at this early stage to ensure that these technologies are developed and deployed in an ethical manner.

Konrad-Adenauer-Stiftung | Digital Futures Lab

Introduction

Governments around the world are recognizing the potential of artificial intelligence (AI) to achieve economic goals and other advantages, as seen by the proliferation of national AI strategies around the world.[1] Countries in South Asia (India, Bangladesh, Bhutan, Nepal, Maldives, Afghanistan Pakistan, and Sri Lanka) have demonstrated varying levels of government readiness to harness AI, as shown by their rankings in Oxford Insights' Government AI Readiness Index, which ranks countries in answer to the question: "How ready is a given government to implement AI in the delivery of public services to their citizens?" [2]

The ranks of South Asian nations are depicted in the table below:

Country	Rank
India	40
Maldives	84
Sri Lanka	90
Bhutan	108
Pakistan	117
Bangladesh	123
Nepal	146
Afghanistan	164

1 Tim Dutton, "*AI Policy 101: An Introduction to the 10 Key Aspects of AI Policy*" (Medium, July 5, 2018) https://medium.com/politics-ai/ai-policy-101-what-you-need-to-know-about-ai-policy-163a2bd68d65; Tim Dutton, "An Overview of National AI Strategies" (Medium, June 29, 2018) https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

2 Oxford Insights and IDRC, "*Government AI Readiness Index 2020.*" (Oxford Insights, 2020). https://static1.squarespace.com/static/58b2e92c1e5b6c828058484e/t/5f7747f29ca3c20ecb598f7c/1601653137399/AI+Readiness+Report.pdf

172 countries are ranked in total. At 40, India is the clear leader in the rankings in South Asia. However, the eight nations also display very wide disparities in rankings throughout the region, from 40 to 164, just three places from the bottom. The 2020 edition also introduces a "Responsible AI Sub-Index", ranking countries on inclusivity, accountability, transparency, and privacy. The index, however, is limited as it ranks only 34 countries in total, although the hope of expanding it in future is expressed. India is the only South Asian country ranked, and is placed 32 out of 34.

In South Asia, therefore, India is clearly the most dominant country in government AI readiness, with other nations lagging behind and somewhat in its shadow. This chapter focuses on the status of one of these nations—Sri Lanka, ranked third in South Asia and 50 places behind India. The aim of this chapter is to embody a perspective from a nation where AI development and the discourse around data and AI governance is still at a very early stage and illustrate some of the opportunities and challenges that are at play. In the early stages of emerging technology, norm-setting is vital to ensure that such technologies are developed in a manner that is ethical and deployed for ethical ends. Anticipating potential pitfalls and looking for ways to address them will be important to engaging issues in AI ethics such as algorithmic bias, accountability, data protection and privacy. The chapter proceeds as follows: First, the landscape and context of AI in Sri Lanka are discussed. Secondly, policy initiatives taken by the government of Sri Lanka concerning AI and data governance are considered. Finally, the chapter discusses gaps in norm-setting in current approaches to AI governance, and what could be done next.

Landscape of AI and Data

AI In The Private Sector and Academia

As is often the case in the early stages of emerging technology, the development of AI technologies and the discourse around its uses and effects appear to be most dominant among the private sector and academia in Sri Lanka at present. For instance, the industry group, the Sri Lanka Association for Software and Services Companies (SLASSCOM), hosts an Artificial Intelligence Centre of Excellence (AICx). The stated aim of the Centre is to "serve as a platform for those who are advocates in AI and passionately involved in different focus areas to come together and help drive the AI agenda for Sri Lanka".[3] The telecommunications industry is a prominent deployer of AI at the moment. The mobile operator Dialog has deployed an AI-based voice assistant,[4] as has the operator SLT-MOBITEL, which launched an AI-

3 "*AI Centre of Excellence (AICx)*," SLASSCOM, accessed October 3, 2021, https://slasscom.lk/ai-centre-of-excellence-aicx/

4 "*Dialog AI Based Voice Assistant*," Dialog, accessed February 21, 2022 https://www.dialog.lk/dialog-ai-based-voice-assistant

powered Facebook chatbot to answer customers' inquiries.[5] In 2021, SLT-MOBITEL also sponsored a hackathon ("Hack: AI 2021 Hackathon") for school and university level students to develop solutions that would address the Sustainable Development Goals (SDGs).[6] Conglomerates such as John Keells Holdings have explored AI use cases to perform fraud analytics and customer profiling.[7] There has also been talk of using AI in the banking sector. In January 2021, the then Governor of the Central Bank stated that the Central Bank would look into AI-powered supervisory technology and regulatory technology for supervising banks off-site.[8] The bank Sampath Bank PLC states that it hosts an AI-powered customer service banking robot (operational in Sinhala, Tamil, and English languages), which can perform functions such as cash withdrawal and queries about account balances.[9]

Sri Lankan academics and researchers working in the fields of data science, machine learning, and AI have produced research on a variety of topics, including using natural language processing,[10] remote sensing,[11] and looking at how human mobility affects disease spread.[12] There have also been collaborations between universities and the private sector. For example, it was reported in June 2021 that the telecommunications company Dialog Axiata PLC had established "5G Innovation Centers" in the Engineering Departments of several Sri Lankan universities.[13] This

5 "*SLT-MOBITEL's BizChat Introduces FB Chatbot to Support SMEs and Micro Businesses Transform Business Operations*," SLT, accessed February 21, 2022. https://www.slt.lk/en/content/slt-mobitel%E2%80%99s-bizchat-introduces-fb-chatbot-support-smes-and-micro-businesses-transform

6 "*SLT-MOBITEL sponsors Hack:AI 2021 to unlock power of Artificial Intelligence and Machine learning to solve real world problems*," SLT, accessed February 21, 2022. https://www.slt.lk/en/content/slt-mobitel-sponsors-hackai-2021-unlock-power-artificial-intelligence-and-machine-learning

7 "*Future of Innovation: Is Artificial Intelligence the Game Changer?*" Echelon, August 17, 2021, https://www.echelon.lk/future-of-innovation-is-artificial-intelligence-the-game-changer/

8 Mahadiya Hamza, "*Sri Lanka Eyes Artificial Intelligence for Bank Regulation: CB Governor,*" EconomyNext, January 16, 2021, https://economynext.com/sri-lanka-eyes-artificial-intelligence-for-bank-regulation-cb-governor-77897/

9 "*Banking Robot: The First Ever Robot in Sri Lanka,*" Sampath Bank, accessed February 21, 2022, https://www.sampath.lk/en/personal/electronic-banking/banking-robot

10 For example: Buddi Gamage, Randil Pushpananda, Ruvan Weerasinghe, and Thilini Nadungodage, "*Usage of Combinational Acoustic Models (DNN-HMM and SGMM) and Identifying the Impact of Language Models in Sinhala Speech Recognition,*" 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), https://www.researchgate.net/publication/350077618_Usage_of_Combinational_Acoustic_Models_DNN-HMM_and_SGMM_and_Identifying_the_Impact_of_Language_Models_in_Sinhala_Speech_Recognition

11 For example – I. P. Senanayake, W.D.D.P. Welivitiya, P.M. Nadeeka, "*Urban green spaces analysis for development planning in Colombo, Sri Lanka, utilizing THEOS satellite imagery—A remote sensing and GIS approach*," Urban Forestry & Urban Greening, 12, no. 3 (2013): 307-314.

12 For example - K.G.S. Dharmawardana, J.N. Lokuge, P.S.B. Dassanayake, M.L. Sirisena, M.L. Fernando, A.S. Perera, and S. Lokanathan, "*Predictive model for the dengue incidences in Sri Lanka using mobile network big data,*" 2017 IEEE International Conference on Industrial and Information Systems (ICIIS),(2017): 1-6, https://ieeexplore.ieee.org/document/8300381

13 "*Dialog Axiata establishes 5G Innovation Centers at leading universities,*" Colombo Page, July 9, 2021. http://www.colombopage.com/archive_21A/Jul09_1625812204CH.php

initiative was taken in collaboration with the University Grants Commission (UGC).[14] The news article reported:

> These Innovation Centers aim to empower students to be interdisciplinary, computational, and entrepreneurial by giving students first-hand experience with 5G connectivity and many of its cutting-edge use cases. This includes artificial intelligence, machine learning, computer vision, blockchain, robotics, the internet of things (IoT) and related emerging technologies that will fuel national innovation and economic diversification.[15]

Mobile operators have also shared anonymized call detail records (CDRs) with big data researchers in Sri Lanka to generate insights on population movements in the city of Colombo (the commercial capital), which can be used to inform urban and transport planning.[16]

Universities are also increasingly offering courses and specializations on AI, machine learning, and data science.[17] However, there still remain challenges to AI research in Sri Lanka. A major obstacle is the lack of easily available, high-quality data. Some academics point out that the lack of Sri Lanka-specific data means that it is difficult for students to research Sri Lanka-specific applications, and they often use public datasets from other national contexts.[18] Sri Lanka does have an Open Data Portal. However, the portal has significant limitations, which will be discussed further later on. Others observe that the distinct separation of disciplines in the Sri Lankan university system makes it harder to engage in interdisciplinary work on AI and the effects of AI.[19] This lack of interdisciplinarity could be a barrier to exploring ethical issues around AI and treating it as a socio-technical system. A mainly technical focus that does not engage with disciplines such as sociology, political science, philosophy etc, means that the societal impacts of AI on the people on whom the technology is deployed will be underexplored.

14 Ibid.

15 Ibid.

16 Maaroof, Abbas, "*Big Data and the 2030 Agenda for Sustainable Development,*" accessed February 21, 2022, https://www.unescap.org/sites/default/files/1_Big%20Data%202030%20Agenda_stock-taking%20report_25.01.16.pdf

17 Ramathi Bandaranayake and Viren Dias. "*Towards a Realistic AI Policy for Sri Lanka: Discussion Paper,*" LIRNEasia (2022). https://lirneasia.net/wp-content/uploads/2022/01/LIRNEasia-Towards-a-Realistic-AI-Policy-for-Sri-Lanka.pdf

18 Ibid.

19 Ibid.

Data and the COVID-19 Pandemic

As was the case in many nations, the COVID-19 pandemic spurred a flurry of initiatives to use data for pandemic control purposes, such as contact tracing and enforcing quarantine. Sources of data in addition to traditional pen-and-paper records that were used included call records, ride-hailing app data, and the "Stay Safe" check-in application.[20] However, these measures have given rise to privacy concerns. The collection of personally identifiable information spurred calls for data protection regulation.[21] In particular, the Stay Safe application led to concerns around privacy and identity theft, with some expressing the worry that the confidentiality of the data would not be ensured.[22] Indeed, cybersecurity concerns about the app emerged when it transpired that a person's information submitted to the application could be checked using an API (Application Programming Interface) call, although the ICT Agency of Sri Lanka (ICTA) later said that security issues had been addressed.[23] Furthermore, others have expressed discomfort with contact tracing methods that require giving personal details such as phone numbers, even through manual methods. Special concerns have been raised by women, and some reported facing sex-based harassment after their phone numbers were taken as part of contact tracing measures.[24] The pandemic has also spurred the production and deployment of new technologies. For instance, in July 2020, the software company MiHCM announced the release of a facial recognition technology that allows for contactless employee check-in and check-out in workplaces, which would allow for a more hygienic process by removing the need to touch surfaces.[25]

It can be seen, therefore, that there is an emerging AI and data landscape in Sri Lanka across sectors such as telecommunications and banking, for purposes including customer service and customer profiling. There is also a burgeoning academic scene producing research outputs on AI and data. However, discussions around AI and data governance appear to be quite limited so far, although it appears that the COVID-19 pandemic has triggered greater discourse around the need for privacy and data protection. The implications of this will be further discussed in

20 Ramathi Bandaranayake, Pirongrong Ramasoota, Ashwini Natesan, and Arthit Suriyawongkul, "*Health-Related Information and COVID-19: A Study of Sri Lanka and Thailand,*" LIRNEasia, (2021). https://lirneasia.net/2021/05/research-report-health-related-information-and-covid-19/

21 Yoshitha Perera, "*The Rising Need for Data Privacy,*" Daily Mirror, October 21, 2020, https://www.dailymirror.lk/recomended-news/The-rising-need-for-data-privacy/277-198315

22 Kamanthi Wickramasinghe, "'*Stay Safe' Sri Lanka Web Portal: Users worry about data privacy and identity theft*," November 13, 2020, https://www.dailymirror.lk/news-features/Stay-Safe-Sri-Lanka-Web-Portal-Users-worry-about-data-privacy-and-identity-theft/131-199772

23 Neville Lahiru, "*The 'Stay Safe' App and Cybersecurity: Should We Be Worried?*" Roar.lk, December 15, 2020. https://roar.media/english/life/technology/stay-safe-cybersecurity-should-we-be-worried

24 Muqadassa Wahid, "*Contact Tracing Raises Privacy Concerns*," Daily Mirror, November 12, 2020. https://www.dailymirror.lk/news-features/Contact-tracing-raises-privacy-concerns/131-199697

25 "*MiHCM announces Facial Recognition capability for Workforce Attendance Authentication,*" Ada Derana, July 21, 2020. http://bizenglish.adaderana.lk/mihcm-announces-facial-recognition-capability-for-workforce-attendance-authentication/

the third section of this chapter. For now, we turn to several policy initiatives by the government of Sri Lanka, concerning data governance.

Policy Initiatives on Data Governance in Sri Lanka

Sri Lanka has taken several policies and regulatory initiatives regarding data. The most prominent of these is the Personal Data Protection Act, which was passed in March 2022,[26] making Sri Lanka the first South Asian country to enact legislation on data protection.[27] However, several other policies are of note, including the draft National Data Sharing Policy [28] and the Sri Lanka Government Information Classification Framework.[29] Sector-specific policies and guidelines have also been introduced for the health sector, with regard to managing health information and health data. These are the National Policy on Health Information (2017)[30] and the National Digital Health Guidelines and Standards [NDHGS] (2020).[31] This reflects a growing acknowledgement of the need for standards and guidance in governing personal data, especially particularly sensitive forms of personal data such as health data. The principles of the National Policy on Health Information include safeguarding confidentiality and privacy. Similarly, the NDGHS includes guidance on maintaining the privacy of personally identifiable data.

Sri Lanka has yet to release a policy or strategy on AI. In 2019, SLASSCOM released a draft policy framework for the promotion of AI in Sri Lanka,[32] however the draft framework was not formally adopted by the cabinet, and there has been no update on this process since the launch of the draft. In February 2022, it was reported that Sri Lanka would receive a grant from India for a "Unitary Digital Identity Framework" which would use biometric data to verify personal identity, supposedly inspired by

26 Personal Data Protection Act, No. 9 of 2022. https://www.parliament.lk/uploads/acts/gbills/english/6242.pdf

27 Rohan Samarajiva, "*Personal Data Protection Act passed: What will it mean?*" Daily FT, March 22, 2022, https://www.ft.lk/columns/Personal-Data-Protection-Act-passed-What-will-it-mean/4-732307

28 "*National Data Sharing Policy (Draft),*" Information and Communication Technology Agency of Sri Lanka," accessed July 31, 2021, http://data.gov.lk/national-data-sharing-policy-draft

29 "*Sri Lanka Government Information Classification Framework (SLGICF),*" accessed 31 July 2021. https://www.gov.lk/elaws/wordpress/wp-content/uploads/2015/08/Information_Classification_FW_Report-v3-1.pdf

30 "*The National Policy on Health Information,*" Ministry of Health, Nutrition, and Indigenous Medicine, accessed July 31, 2021, http://www.health.gov.lk/moh_final/english/public/elfinder/files/publications/publishpolicy/NationalPolicyonHealthInformation.pdf

31 "*National Digital Health Guidelines and Standards [NDHGS] 2.0,*" Ministry of Health, accessed February 21, 2022. http://www.health.gov.lk/moh_final/english/public/elfinder/files/publications/list_publi/NDHGS%20v2.pdf

32 Hiyal Biyagamage, "*SLASSCOM launches Sri Lanka's first AI policy framework,*" Daily FT, June 27, 2019, https://www.ft.lk/Front-Page/SLASSCOM-launches-Sri-Lanka-s-first-AI-policy-framework/44-680805

India's Aadhar system.[33] This illustrates regional and cross-country influences in data policy and governance. At the time of writing, however, there has been no further update on this.

Personal Data Protection Act

Some parts of the recently passed Personal Data Protection Act have relevance for AI. For example, Section 18 has regulations on automated processing, which would be relevant for an entity which seeks to use AI or algorithmic decision-making in its operations:

SECTION 18 (1)

> Subject to section 19, every data subject shall have the right to request a controller to review a decision of such controller based solely on automated processing, which has created or which is likely to create an irreversible and continuous impact on the rights and freedoms of the data subject under any written law.

SECTION 18 (2)

notes the conditions under which subsection (1) does not apply:

> where a decision of a controller, based on automated processing is —
>
> (a) authorized by any written law, which a controller is subject to;
>
> (b) authorized in a manner determined by the Authority;
>
> (c) based on the consent of the data subject; or
>
> (d) necessary for entering into or performance of a contract between the data subject and the controller, and the controller shall comply with such measures and applicable criteria as may be specified by the Authority by rules made in that behalf to safeguard the rights and freedoms of the data subject:
>
> **Provided however, the requirement under paragraph (d) shall not apply to special categories of personal data.**

33 Meera Srinivasan, "*India to help Sri Lanka launch its version of Aadhaar,*" The Hindu, February 9, 2022, https://www.thehindu.com/news/international/india-to-help-sri-lanka-launch-its-version-of-aadhaar/article38395719.ece

Draft National Data Sharing Policy

The draft policy states that it will apply "to all data created, generated, collected or archived by the Government of Sri Lanka through its associated departments/ ministries/ agencies using public funds. The data may be in electronic form or in form of manual records." It aims to facilitate smooth data sharing between government departments and the general public. It hopes to advance principles of open government and open data. The policy also classifies three levels of access to data: open access (any individual may freely access the data without any prior registration or process), authorized access (data may be accessed only through a process of registration or authorization by specific government departments), restricted access (access granted only to specific authorized organizations or individuals on a need to know basis).

Sri Lanka Government Information Classification Framework

This framework provides a classification of government information by levels of security. According to the framework: "This could be used for the purpose of government data sharing in accordance with the Right To Information (RTI)." The levels of security classification given in the framework are:

- **Unclassified:** Information yet to be classified
- **Public:** "Any information which is easily available to the public, Government employees, organizations, regulators, project managers, support staff and contractors including information deemed public by legislation or through a policy of routine disclosure- this type of information requires minimal or no protection from disclosure"
- **Limited** sharing: "Information is security classified as "Limited Sharing" when compromise of information may lead to minor probability of causing limited damage to Sri Lankan Government, commercial entities or members of the public. Unauthorized disclosure of this information will cause negligible or no damage to internal security, Sri Lankan forces or Sri Lanka's foreign relations"
- **Confidential:** "compromise of information may lead to a high probability of causing damage to national security, internal stability, national infrastructure, forces, commercial entities or members of the public"
- **Secret:** "compromised could cause serious damage to national security, Government, nationally important economic and commercial interests or threaten life; it could also

> raise international tension and seriously damage relations with other governments, shut down or substantially disrupt significant national infrastructure and seriously damage the internal stability of Sri Lanka or other countries"

The draft National Data Sharing Policy and the Sri Lanka Government Information Classification Framework are of note because they could impact the kind of government data that can be accessed by AI and data science researchers. The Sri Lankan government has taken some steps towards Open Data, which we now turn to.

Open Data Portal

Open data portals can be valuable sources of freely, easily accessible data. These can be of great benefit to researchers. As of now, Sri Lanka's Open Data Portal contains a total of 136 datasets. The table below shows the number of datasets classified by their types/themes as given in the portal:

Eight principles are listed on the Open Data Portal for government data to be considered open (quoted below),[34] based on the principles of open government data published on opengovdata.org:

Table 3:

Type	Number of Datasets
Agriculture and Livelihood	43
Tourism and Leisure	1
Travel	1
Information Technology and Cyber Security	21
Demography	19
Economic	17
Employment and Skills	5
Industry and Investments	4
Infrastructure	8
National Security and Safety	6
Transport	11

34 "*The 8 Principles of Open Government Data,*" Information and Communication Technology Agency of Sri Lanka," accessed July 31, 2021. http://www.data.gov.lk/8-principles-open-government-data

- **Data** must be complete
- **Data** must be timely
- **Data** must be machine processable
- **Data** formats must be non-proprietary
- **Data** must be primary
- **Data** must be accessible
- **Access** must be non-discriminatory
- **Data** must be license-free

Sri Lanka has made commitments to open data under the Open Government Partnership (OGP). This includes increasing the number of datasets available in the Open Data Portal.[35] However, there are several issues with the portal. Dias and Bandaranayake (2021) have discussed some of its shortcomings.[36] For one, Sri Lanka intended to increase the number of datasets on the portal to 200 (by end of 2016) and then 500 (by July 2018) in the OGP's 2016-18 action plan cycle. However, this milestone was not met by the end of the action plan cycle. Furthermore, several datasets on the portal have been observed to be not in machine-readable formats, and collection methodologies and variable descriptions have not been sufficiently documented. Dias and Bandaranayake also identified three datasets of beekeepers from three districts in Sri Lanka which contained personally identifiable information such as telephone numbers, addresses, and names. They note that regardless of whether consent was obtained, such detailed personal information would be of little value to researchers using the dataset. Hence there appears to be little rationale for exposing such personal details on the portal. In 2017, Máchová and Lnénicka published an evaluation of the quality of national open data portals. Sri Lanka ranked 47 out of 67 countries ranked.[37] Among the other South Asian countries ranked, India is ranked 2, Nepal is ranked 20, and Pakistan 43. This shows that relative to other South Asian countries, Sri Lanka's Open Data Portal still lags behind.

35 "*Promote the Open Data Concept and Delivering the Benefits to Citizens Through ICT (LK0006),*" Open Government Partnership, accessed July 31, 2021, https://www.opengovpartnership.org/members/sri-lanka/commitments/LK0006/; "*Open Data (LK0029),*" Open Government Partnership, accessed July 31, 2021, https://www.opengovpartnership.org/members/sri-lanka/commitments/LK0029/

36 Viren Dias and Ramathi Bandaranayake, "*Sri Lanka's Open Data Portal: Current Status and Opportunities for Improvement*," LIRNEasia (2021). https://lirneasia.net/wp-content/uploads/2021/12/LIRNEasia-Sri-Lankas-Open-Data-Portal.pdf

37 Renata Máchová and Martin Lnénicka, "*Evaluating the Quality of Open Data Portals on the National Level,*" Journal of theoretical and applied electronic commerce research, 12, no. 1(2017). http://dx.doi.org/10.4067/S0718-18762017000100003

AI Governance: Exploring Ethical Questions

So far in this chapter, we have considered some aspects of the current landscape of AI and data in Sri Lanka. However, the discourse on the governance of these emerging technologies is still very limited. Below, several ethical questions related to AI and data are discussed, along with their relation to the Sri Lanka context.

Privacy and Surveillance

Discourse around privacy has certainly gathered steam with the passage of the Personal Data Protection Act. The draft National Data Sharing Policy, Government Information Classification Framework, and the two health policy guidelines also show sensitivity to the need to protect personal data. However, the COVID-19 pandemic has highlighted some of the risks to privacy. On a positive note, the pandemic has prompted discourse on some of the risks entailed in the collection of personal data. However, with the introduction of check-in applications and facial recognition technologies as part of the pandemic response, there is a danger that such measures may become normalized if they are regarded as essential for a return to normal life. This could lead to "surveillance creep", where the intensive monitoring and collection of personal data continues even after the emergency that triggered the data collection passes.[38] Therefore, while this initial discourse is promising, the use of such technologies should continue to be monitored to ensure they are not overused.

Furthermore, if Sri Lanka is to proceed with the "Unitary Digital Identity Framework" apparently based on Aadhar, it would be prudent to keep in mind criticisms of Aadhar. For instance, it has been pointed out that the Aadhar system's reliance on biometrics excludes those who are experiencing poverty, and who may lack iris scans that can be used because of malnutrition, or those who lack fingerprints due to the effects of hard labour.[39]

The passage of the Personal Data Protection Act has also not been without debate. An opposition legislator argued that the proposed Data Protection Authority was not independent enough.[40] Transparency International Sri Lanka (TISL) echoed this concern, stating that the proposed Authority lacked adequate safeguards to prevent political interference.[41] TISL also expressed concern that "journalistic purposes" for

38 For example, see Mike Giglio, "*Would you sacrifice your privacy to get out of quarantine?*" The Atlantic, April 22, 2020, https://www.theatlantic.com/politics/archive/2020/04/coronavirus-pandemic-privacy-civil-liberties-911/609172/

39 Linnet Taylor, "*What is data justice? The case for connecting digital rights and freedoms globally,*" Big Data & Society (2017). https://journals.sagepub.com/doi/full/10.1177/2053951717736335

40 "*Sri Lanka parliament passes data protection act amid privacy concerns,*" Economy Next, March 10, 2022. https://economynext.com/sri-lanka-parliament-passes-data-protection-act-amid-privacy-concerns-91476/

41 Ibid.

using personal data are not recognized, which could pose barriers to reporting.[42] Others opined that the requirement for purpose specification at the time data is being collected could limit innovation in AI, since it would be harder to repurpose the data for machine learning applications.[43]

Bias

With the intended use of AI technologies for decision-making in applications such as customer profiling (as mentioned previously), the question of algorithmic bias comes into play. One source of bias is the lack of representivity in datasets. Since many policies and initiatives (such as the Open Data Portal) have focused on access to data, it is worth considering the questions of bias and representivity here. Bandaranayake and Dias (2022) illustrate this issue by using the following example: if a dataset of de-identified call detail records (CDRs) from Sri Lanka were used as a proxy for human mobility, the dataset would exclude the population that does not use mobile phones. Furthermore, Sri Lanka has a gender gap in mobile phone ownership, as well as a rural-urban divide. Hence, these populations would be underrepresented in the dataset.[44] Therefore, researchers would need to be sensitive to these imbalances in the data. Likewise, policymakers who use big data research as part of their evidence base need to be aware of these limitations. Private sector entities who are considering using big data sources to create consumer profiles and for other applications should be aware of this as well.

Issues of the potential for discrimination have been identified with regard to open data as well. It has also been argued that even if individual data is protected, this may not prevent the targeting of people based on certain group characteristics—"Open Data is more likely to treat types (of customers, users, citizens, demographic population, etc) rather than tokens (you, Alice, me…), and hence groups rather than individuals. But re-identifiable groups are ipso facto targetable groups."[45] It has also been contended that potential problems with open data include "the embedding of social privilege in datasets as the data is constructed, [and] the differential capabilities of data users."[46] This has the potential to lead to certain forms of discrimination based on group characteristics.

42 Ibid.

43 Rohan Samarajiva, "*Personal Data Protection Act passed: What will it mean?*" Daily FT, March 22, 2022, https://www.ft.lk/columns/Personal-Data-Protection-Act-passed-What-will-it-mean/4-732307

44 Ramathi Bandaranayake and Viren Dias. "*Towards a Realistic AI Policy for Sri Lanka: Discussion Paper,*" LIRNEasia (2022). https://lirneasia.net/wp-content/uploads/2022/01/LIRNEasia-Towards-a-Realistic-AI-Policy-for-Sri-Lanka.pdf

45 Luciano Floridi, "*Open Data, Data Protection, and Group Privacy,*" Philos. Technol. 27 (2014) https://link.springer.com/content/pdf/10.1007/s13347-014-0157-8.pdf

46 Jeffrey Alan Johnson, "*From open data to information justice,*" Ethics and Information Technology, 16 (2014) https://link.springer.com/article/10.1007%252Fs10676-014-9351-8

Explainability and Accountability

Two of the key questions in AI ethics are explainability and accountability. When it comes to algorithmic decision-making using machine learning, the "black box" makes it tricky to explain how a given decision was arrived at. Similarly, there are questions of who should be held to account for an algorithmic decision. The programmer? The person who deployed it? Can the algorithm itself be held responsible? These questions take on greater significance when the decisions are made in high-stakes scenarios, for example concerning health or finance. There has been some interest in Sri Lanka in developing AI solutions for banking applications. When developing such applications, it would be useful to bear in mind existing debates around AI and algorithmic decision-making ethics in finance. For example, a prominent topic of discussion has been the use of AI to create alternate credit scores to try and increase financial inclusion for those who face trouble borrowing due to a lack of existing credit history.[47] However, ethical complications have been pointed out, such as the possibility of algorithmic bias against certain populations.[48] As the banking sector looks to further various applications of AI in Sri Lanka, it would be wise to keep these lessons in mind.

Furthermore, recall Section 18 (1) of the Personal Data Protection Act:

> **Subject to section 19, every data subject shall have the right to request a controller to review a decision of such controller based solely on automated processing, which has created or which is likely to create an irreversible and continuous impact on the rights and freedoms of the data subject under any written law.**

While it would be unrealistic at the moment to expect full explainability and an opening of the "black box", those creating and deploying AI solutions would need to bear this requirement in mind, and find ways of being accountable for decisions made using automated processing.

47 See for example Vira Widiyasari and Herman Widjaja, "*This new approach to credit scoring is accelerating financial inclusion in emerging economies*," World Economic Forum, January 20, 2021, https://www.weforum.org/agenda/2021/01/this-new-approach-to-credit-scoring-is-accelerating-financial-inclusion/

48 See for example Hicham Sadok, Fadi Sakka, and Mohammed El Hadi El Maknouzi, "*Artificial intelligence and bank credit analysis: A review*," Cogent Economics & Finance, 10, no. 1 (2022). https://www.tandfonline.com/doi/full/10.1080/23322039.2021.2023262

Going Forward: How should norms be set?

Given that these ethical issues have been identified, how should norms be set?

> The fact that AI development and deployment are still at an early stage provides a good opportunity to set norms early, and ensure that discourse about the ethics and social impacts of AI and data go hand in hand with technological advances.

Firstly, awareness about ethical issues in data and AI needs to be increased. This could include integrating modules on ethics and social impacts into university curricula on AI, machine learning and data science, so that students learn how to think about the development of these systems holistically instead of as purely technical phenomena. Awareness of issues such as privacy, bias, explainability and accountability should also be increased among the private sector. The private sector itself can become a norm-setter through its own practices, so it is important to ensure that these ethical principles are part of this conversation. Policymakers could have a role to play here as well. For example, Singapore released its Model AI Governance Framework in 2019, with a second edition being released in 2020. This framework provides high-level guidance on ethical principles for organizations developing and deploying AI, discussing issues such as fairness, transparency, and the need for stakeholder engagement, among others.[49] Policymakers in Sri Lanka could consider a similar initiative. Civil society can also play a role in tracking real-world implementations of ethical principles, highlighting good practices, and pointing out areas for improvement.

Secondly, it was mentioned previously in this chapter that AI should be treated as a socio-technical system. This means that the impacts of the system on the community in which it is deployed should be considered. A variety of tools and frameworks have been released that could help address this. For example, researchers at AI Now have proposed a framework for public agencies to assess the effects of automated decision systems.[50] Similarly, the government of Canada has released an algorithmic

49 Infocomm Media Development Authority, Singapore, "*Model Artificial Intelligence Governance Framework: Second Edition*," (2020), accessed February 21, 2022, https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

50 Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker, "*Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*," AI Now, (2018), https://ainowinstitute.org/aiareport2018.pdf

impact assessment tool.[51] Creating and deploying similar tools in a manner sensitive to the Sri Lankan context would help researchers, private sector entities, and policymakers better understand the social impacts of AI technologies, and proactively address any concerns from the people the technology is being deployed on.

Finally, it is necessary to address issues with regard to data. Difficulty in accessing quality data is one of the major obstacles to AI research and development and can be a source of bias. The Sri Lankan government should make efforts to revamp the Open Data Portal by increasing the number of datasets and updating them. Government departments should be incentivized to contribute more datasets to the portal. Data governance initiatives that are in draft form, such as the draft National Data Sharing Policy, should be discussed, formally adopted, and actioned as soon as possible in order to provide oversight into the uses of data. At the time of writing, the Personal Data Protection Act has just been passed, so it remains to be seen what its effects will be. Data sharing between the private sector and researchers should also be further encouraged.

Overall, this chapter has detailed some of the emerging discourses around AI, data, and the governance of these technologies in Sri Lanka, and has illustrated what this discourse looks like in the early stages in a relatively resource-constrained country. The nascent stage of this discourse provides an ideal opportunity to influence norms to ensure that the benefits of AI are maximized, and the harms minimized.

51 "*Algorithmic Impact Assessment Tool*," Government of Canada, accessed February 21, 2022. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

To What Extent Does Malaysia's National Fourth Industrial Revolution Policy Address AI Security Risks?

Digital Futures Lab | Konrad-Adenauer-Stiftung

02

CHAPTER 02

To What Extent Does Malaysia's National Fourth Industrial Revolution Policy Address AI Security Risks?

JUN-E TAN

Abstract

The National Fourth Industrial Revolution (4IR) Policy was launched in July 2021 as a guiding document for Malaysia's direction in maximizing growth opportunities and mitigating potential risks arising from 4IR technologies. This chapter explores the policy to examine the extent to which Artificial Intelligence (AI) security risks are addressed, using the AI Security Map by Newman (2019) as a framework. In the policy, 4IR technologies including AI are seen through a techno-utopian lens, therefore its focus centres on rapid adoption rather than regulation and resilience. It is found that most of the policy initiatives focus on economic security and capacity building for the state, in order to keep up with the developmental race. Other areas of AI security such as the risks of unintended consequences or unsafe outcomes of AI, or risks of AI being used for malicious purposes, receive much less attention. However, as the N4IRP is still in its nascent stages of implementation, there is still room for its cross-ministerial governance structure to work on providing safeguards across different domains and sectors to achieve holistic and sustainable development.

Introduction

Artificial intelligence (AI) and the Fourth Industrial Revolution (4IR) have become the next big thing in the developmental race, as countries attempt to harness technology to get ahead, or at least to not be left behind. The Fourth Industrial Revolution describes a transformation in the ways we live, work, and communicate through the application of a range of technologies fusing the physical, digital, and biological worlds.[1] As much as some breakthroughs in technology had powered previous industrial revolutions—such as mechanization with the steam engine, mass production with electricity, and computerization with the semiconductor—AI (defined in this context as algorithms generating algorithms[2]), is one of the foundational technologies that will open up a new era of industrialization.

The National Fourth Industrial Revolution Policy[3] (N4IRP) of Malaysia was launched in July 2021, with the aim of "driving coherence in transforming the socioeconomic development of the country through ethical use of 4IR technologies". The key foci of the policy are on maximizing growth opportunities and on mitigating potential risks arising from 4IR. The policy includes AI as one of five foundational 4IR focal areas, the others being the Internet of Things (IoT), blockchain, cloud computing and big data analytics, and advanced materials and technologies. Among these technologies, AI "is expected to create the most impact", and is considered "the 'electricity' of the 4IR".[4]

Within this chapter, the lens of AI security is used to scrutinize the pathway towards 4IR in Malaysia. AI as a transformative technology has the potential to bring not only societal benefits, but also harms to society that may negate developmental gains. Already in other parts of the world, we see some of these harms in the form of unintentional consequences such as amplified systemic biases resulting in the marginalized being further marginalized,[5] or weaponized AI which efficiently surveil entire populations[6] or carry out automated cyberattacks.[7] The evolution of the technology outpaces the speed in which legal and regulatory safeguards are put in

1 Klaus Schwab, The Fourth Industrial Revolution (Geneva: World Economic Forum, 2016).

2 Internet Society, "*Artificial Intelligence and Machine Learning: Policy Paper*" (Internet Society, April 2017), https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/.

3 Government of Malaysia, "*National Fourth Industrial Revolution (4IR) Policy*" (Economic Planning Unit, Prime Minister's Department of Malaysia, July 2021), https://www.epu.gov.my/sites/default/files/2021-07/National-4IR-Policy.pdf.

4 N4IRP, p59

5 Ed Pilkington, "*Digital Dystopia: How Algorithms Punish the Poor*", The Guardian, 14 October 2019, sec. Technology, https://www.theguardian.com/technology/2019/oct/14/automating-poverty-algorithms-punish-poor.

6 Yael Grauer, "*Surveillance of Uyghurs Detailed in Chinese Police Database*", The Intercept, 29 January 2021, https://theintercept.com/2021/01/29/china-uyghur-muslim-surveillance-police/.

7 Center for Security and Emerging Technology, Micah Musser, and Ashton Garriott, "*Machine Learning and Cybersecurity: Hype and Reality*" (Center for Security and Emerging Technology, June 2021), https://doi.org/10.51593/2020CA004.

place,[8] and also the ability of the average citizen to understand the implications of the technology and how best to protect oneself against possible dangers.

Situating AI technologies within the application contexts of 4IR brings some advantages. On one hand, the perspective of risks and harms is anchored in applications and implications instead of focusing only on technological limitations and errors. On the other hand, as 4IR covers a wide range of emerging technologies at various stages of maturity, a focus on AI narrows down the possible risks into a smaller array of known issues, which helps in concretizing problems and imagining solutions. That being the case, even though 4IR technologies would have a larger set of security risks, the concern of this chapter is on the ones that are associated with AI.

Malaysia's N4IRP explicitly acknowledges that there will be potential risks arising from 4IR technologies, and states the government's commitment to address them. The objective of this chapter is therefore to review Malaysia's N4IRP, its goals, and in particular its outlined initiatives, to understand the types of AI-related risks it addresses. The chapter also aims to provide a perspective of technology governance from a developing country's context through delving into Malaysia's priorities in balancing the need to be competitive at an international level, yet protect its citizens' well-being locally.

To what extent does Malaysia's N4IRP address AI security risks? This question requires us to first unpack what AI security risks are, which we will do via the types of potential AI security risks from the AI Security Map proposed by Jessica Newman (2019).[9] We then go through a background of developmental policies of Malaysia to situate the N4IRP, and describe the structure of the policy's content. An analysis is provided on the types of AI security risks covered by the policy and the gaps in risk mitigation. The chapter ends with a discussion on assumptions behind the policy direction, and possible implications on technology governance on the country.

Types of AI Security Risks

What are the security risks of AI? In this section, we explore a framework by Jessica Newman, of the Center for Long-Term Cybersecurity, which lists out twenty types of such risks, organized into digital/physical, political, economic, and social domains (see Table 1). Within her 2019 paper, *Toward AI Security: Global Aspirations for a More Resilient Future*, Newman provides comprehensive examples of the risk areas, and also uses the AI Security Map to analyse national AI strategies and policy responses

8 Gary Marchant, "*Governance of Emerging Technologies as a Wicked Problem*", Vanderbilt Law Review 73, no. 6 (1 December 2020): 1861.

9 Jessica Cussins Newman, "*Toward AI Security: Global Aspirations for a More Resilient Future*", CLTC White Paper Series (Berkeley: Centre for Long-term Cybersecurity, February 2019), https://cltc.berkeley.edu/wp-content/uploads/2019/02/CLTC_Cussins_Toward_AI_Security.pdf.

of ten countries to determine their preparedness in handling AI security threats and opportunities.

Table 1: AI Security Map (Newman, 2019)

AI SECURITY DOMAINS			
Digital/Physical	**Political**	**Economic**	**Social**
Reliable, value-aligned AI systems	Protection from disinformation and manipulation	Mitigation of labour displacement	Transparency and accountability
AI systems that are robust against attack	Government expertise in AI and digital infrastructure	Promotion of AI research and development	Privacy and data rights
Protection from the malicious use of AI and automated cyberattacks	Geopolitical strategy and international collaboration	Updated training and education resources	Ethics, fairness, justice, dignity
Secure convergence / integration of AI with other technologies (bio, nuclear, etc.)	Checks against surveillance, control, and abuse of power	Reduced inequalities	Human rights
Responsible and ethical use of AI in warfare and the military	Private-public partnerships and collaboration	Support for small businesses and market competition	Sustainability and ecology

Newman defines AI security "as the robustness and resiliency of AI systems, as well as the social, political, and economic systems with which AI interacts", and looks beyond the narrow scope of national security to cover a more comprehensive landscape of security issues. The AI Security Map was chosen as a point of reference because it provides a comprehensive (but, as Newman emphasizes, not exhaustive) overview of the breadth of issues that can be included as AI security risks and risk mitigation. The systemic nature of the risks highlighted by the framework is suitable for national-level analyses; Newman goes beyond risks and accountability issues at a technical level that are often focused upon in discussions on AI governance,[10] and looks at a more holistic range of potential harms on society. That Newman has used the framework to conduct analyses on other countries also helps to provide some global context and different national priorities for comparison.

10 Thilo Hagendorff, "*The Ethics of AI Ethics: An Evaluation of Guidelines*", Minds and Machines 30, no. 1 (1 March 2020): 99–120, https://doi.org/10.1007/s11023-020-09517-8.

The digital/physical domain of security risks focuses on various aspects of AI systems design and use that can threaten the security of intertwined digital and physical spaces. The political domain focuses on different actors and their interactions within the AI landscape, between state, market, and society. Relationships between actors reflect power imbalances and priorities that are at times aligned (such as between public and private entities), or at times conflictive (such as government surveillance on populations). The economic domain of AI security risks considers on one hand the impacts of AI technologies on the economy, and on the other, the importance of dedicating resources to drive the technology sector in order to not be left behind. For the social domain, security risk mitigation comes in the form of building in principles, rights, and obligations into AI technologies so that negative impacts on society can be minimized.

This researcher takes the liberty to simplify the 20 security areas into the mitigation of three types of risks: 1) the risks or opportunity costs of not implementing AI, missing out on potential benefits; 2) the risks of unintended consequences or unsafe outcomes of AI; and 3) the risks of AI being used for malicious purposes. We will return to the AI security risks later on, and now introduce Malaysia's N4IRP and its policy landscape related to AI.

Malaysia's National 4IR Policy

Background

The N4IRP was unveiled in early July 2021, as a sister policy to the Malaysia Digital Economy Blueprint[11] (MDEB) which was launched in March 2021. The scope of the MDEB is broader, aiming to "transform Malaysia into a digitally-driven, high income nation, and a regional leader in digital economy", whereas the N4IRP zooms in a little closer into transforming the country's socio-economic development through the use of 4IR technologies, by providing key guiding principles and strategic direction, as well as guidelines to addressing risks.

The MDEB and N4IRP are expressly built to support and enable national development, as their goals are aligned with the objectives of Malaysia's developmental master

11 Government of Malaysia, "*Malaysia Digital Economy Blueprint*" (Economic Planning Unit, Prime Minister's Department of Malaysia, March 2021), https://www.epu.gov.my/sites/default/files/2021-02/malaysia-digital-economy-blueprint.pdf.

plans (the Shared Prosperity Vision 2030[12] and the 12th Malaysia Plan[13] are referenced directly). These two policies are intertwined in that both are administered by the National Digital Economy and 4IR Council which is chaired by the Prime Minister, therefore they share a governance structure. The N4IRP also includes a page on how both policies complement each other. The two policies extend Malaysia's past efforts in developing its digital economy and high-tech ecosystem, notably through the Multimedia Super Corridor (MSC) initiative from the 1990s, to create an IT hub within the country in a version of Silicon Valley. Malaysia's path towards digitalization has been lined with several other policies, such as the National eCommerce Roadmap, the National Industry 4WRD Policy, the National IoT Framework, the National Big Data Analytics (BDA) Framework, the National Fiberisation and Connectivity Plan 2019-2023 (NFCP), and so on.

Malaysia also has policy documents that are focused on AI specifically. There are at least two: the National AI Roadmap (AI-RMap) that was launched in March 2021 by the Ministry of Science, Technology, and Innovation (MOSTI), and the National AI Framework by the Malaysia Digital Economy Corporation (MDEC)[14] which does not appear to have been released publicly.[15] The AI-RMap project was conducted by professors from Universiti Teknologi Malaysia (UTM) and industry experts from the National Tech Association of Malaysia (PIKOM), who were awarded a grant by MOSTI to study and propose paths forward in the area of AI. It was launched in a virtual town hall (because of movement control under the COVID-19 pandemic), introducing the current situation of AI in Malaysia, strategies to diffuse the technology, and proposed national AI projects.[16]

From the AI-RMap website and the few available media reports, it is not readily apparent if the Roadmap is at the stage of being proposed or it is already under implementation.[17] Even though the Roadmap offers specific timelines and action plans between 2021 and 2025, there are very few media reports covering the Roadmap

12 The Shared Prosperity Vision (SPV) 2030 was launched in 2019 by then Prime Minister Mahathir Mohamad. A key aspirational document that is referenced repeatedly in other policies, SPV 2030 provides a longer term direction, with the primary aim of providing a "decent standard of living to all Malaysians by 2030", elaborated within its three objectives: 1) providing development for all; 2) addressing wealth and income disparities; and 3) building a united, prosperous and dignified nation. SPV 2030 continues the tradition of Malaysia's policy formula of growth, distribution, and unity, from previous grand plans such as the New Economic Policy (1971-1990), Vision 2020 (1991-2020), and the New Economic Model (2010-2020).

13 Malaysia has five-year plans which steer the direction of the country's development. The 12th Malaysia Plan covers the period of 2021 to 2025.

14 MDEC is the lead government agency instrumental in developing Malaysia's ecosystem for information and communication technologies and digital economy since the 1990s.

15 There was no launch media article or announcement found within MDEC's database of press releases. However, the framework was referenced within the AI-RMap website with a snapshot of its cover.

16 The contents of the Roadmap are available in https://airmap.my, in the form of slides and also videos of presentations given during the town hall, which happened on March 15, 2021.

17 Attempts were made to reach out to the project leader, with no response.

itself, the virtual town hall event, or its proposed activities.[18] The Roadmap was launched by the Secretary General of MOSTI during the virtual town hall, but there is no mention of the Roadmap in the ministry's website. That it is addressed as a "living document"[19] adds to the tentativeness of the initiative. Erring towards the side of caution, analyses within this chapter focus on N4IRP to indicate Malaysia's priorities and direction when it comes to AI adoption and governance, within a larger context of the Fourth Industrial Revolution.

The Structure of the National 4IR Policy

The N4IRP is published by the Economic Planning Unit of the Prime Minister's Department. Its vision is to harness the power of 4IR technologies to enhance socio-environmental well-being and economic growth. Three missions are outlined: to improve quality of life by leveraging technological advancement, to enhance local capabilities to embrace 4IR across sectors, and to use the technologies to enhance the preservation of ecological integrity. In other words, the N4IRP aims for 4IR technologies, the chief of which is AI,[20] to be used "for good", from social, economic, and environmental points of view. The objectives stated are to seize growth opportunities arising from the 4IR, to create a conducive ecosystem to cope with the 4IR, and to build trust in an inclusive digital society.

The range of technologies covered by the N4IRP is broad, described as new technology that is characterized by "the fusion of physical, digital, and biological worlds, impacting all disciplines, industries and the economy". It covers building capacities in five foundational technologies: 1) artificial intelligence; 2) Internet of Things; 3) blockchain; 4) cloud computing and big data analytics; and 5) advanced materials and technologies, and capabilities in these are expected to be applied across ten key economic sectors[21] and six supporting sectors.[22]

As can be seen in Figure 1, the four *policy thrusts,* or thematic foci of the N4IRP revolve around 1) capacity development and skills training; 2) digital infrastructure development; 3) regulation; and 4) accelerating 4IR technology innovation and adoption. These are broken down into 16 *strategies*, colour coded by "beneficiary groups", which are businesses, government, and society. The 16 strategies are expanded into 32 national initiatives, which have specific timelines assigned to each: initiatives within Phase One to be completed by 2022, Phase Two by 2025, and Phase Three by 2030. For the ten key economic sectors, there are 60 *sectoral initiatives*

18 The most comprehensive report found was one in the Newshub section of Universiti Teknologi Malaysia (https://news.utm.my/2021/07/ahibs-experts-entrusted-for-ai-roadmap-and-talent-development-in-malaysia/). No mention was found in mainstream news media.

19 Malaysia Artifical Intelligence Roadmap. https://airmap.my/ai-roadmap-overview

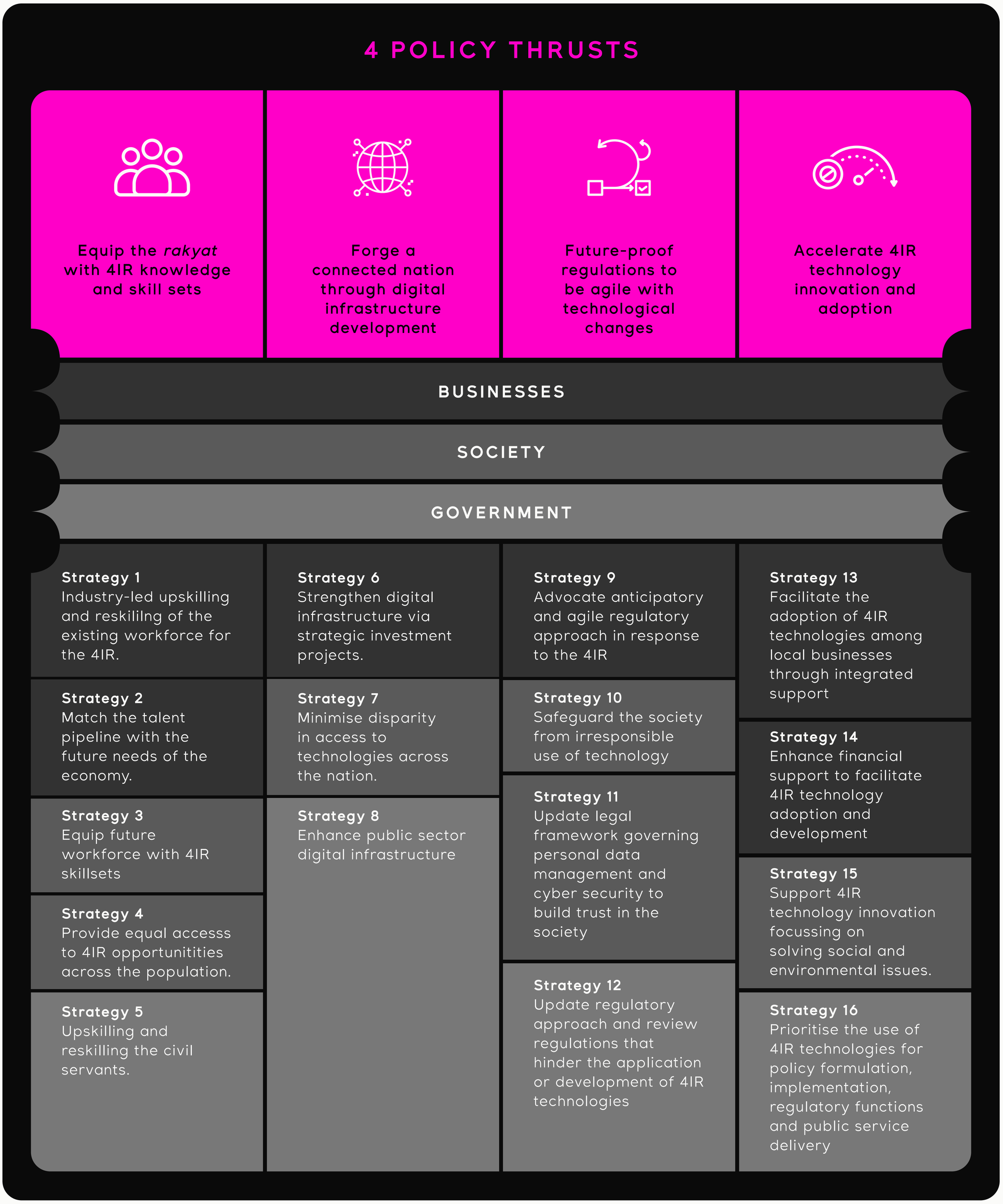
20 Page 59 of the N4IRP

21 Including 1) wholesale and retail trade; 2) transportation and logistics; 3) tourism; 4) finance and insurance; 5) utilities; 6) professional, scientific and technical services; 7) healthcare; 8) education; 9) agriculture; and 10) manufacturing.

22 Including 1) construction; 2) real estate; 3) mining and quarrying; 4) information and communication services; 5) arts, entertainment and recreation services; 6) administrative and support services.

which also align with the four policy thrusts, with some sectoral nuances but mostly adhering to the same themes.

Figure 1: Screenshot of policy thrusts and strategies from the N4IRP



Mapping AI Security Risk Mitigation in the N4IRP

In this section, we provide an analysis of AI security risk mitigation in the N4IRP based on the framework of Newman's AI Security Map. As the scope of the N4IRP covers a wider range of technologies than only AI, the policy may be justifiably vague in some coverage of AI security risks. AI in the policy is also addressed from the angle of 4IR, therefore not all AI security risks within Newman's framework may fit within the context of the policy. For example, "protection from disinformation and manipulation" as listed in the map may not be considered as relevant to the 4th Industrial Revolution. However, there is still merit in the exercise of measuring Malaysia's mitigation of AI security risks according to Newman's framework, as most of the risks listed do still apply under the N4IRP, and we will still be able to identify gaps at the domain level.

To separate the rhetoric from the implementation priorities, emphasis is put on examining the national initiatives outlined under the N4IRP's strategies to be carried out in the next decade. These initiatives are concrete action plans with timelines attached, and represent stated commitment by the government to address certain issues. Table 2 provides an overview of Malaysia's plans to mitigate AI security risks, sorting the security areas into three categories: 1) a clear commitment by the N4IRP to address the issue, based on its inclusion in the planned initiatives; 2) indirect reference or acknowledgment within the policy document, which implies possible action; and 3) no mention of the security risk area, which implies a lower likelihood of the risk being managed. Since the initiatives are not described in detail within the policy, there are some ambiguities which require interpretation and assumptions, which are explained below the table, in the order of priority within the N4IRP.

Table 2:
Malaysia's priorities on AI security risk mitigation

		CLEAR COMMITMENT	INDIRECT REFERENCE/ POSSIBLE INCLUSION	NO MENTION
ECONOMIC	Mitigation of labour displacement	/		
	Promotion of AI research and development	/		
	Updated training and education resources	/		
	Reduced inequalities		/	
	Support for small businesses and market competition	/		
POLITICAL	Protection from disinformation and manipulation			/
	Government expertise in AI and digital infrastructure	/		
	Geopolitical strategy and international collaboration	/		
	Checks against surveillance, control, and abuse of power			/
	Private-public partnerships and collaboration	/		
SOCIAL	Transparency and accountability		/	
	Privacy and data rights	/		
	Ethics, fairness, justice, dignity	/		
	Human rights		/	
	Sustainability and ecology	/		
DIGITAL/PHYSICAL	Reliable, value-aligned AI systems		/	
	AI systems that are robust against attack		/	
	Protection from the malicious use of AI and automated cyberattacks		/	
	Secure convergence / integration of AI with other technologies (bio, nuclear, etc.)			/
	Responsible and ethical use of AI in warfare and the military			/

Economic Domain

Economic security is the highest in priority for the N4IRP. This is unsurprising, given that the document was launched by the Economic Planning Unit, and that AI is framed within the context of 4IR and the digital economy. From the 32 national initiatives, more than half (at least 17) are directly linked to the economy, mostly in supporting the promotion of AI research and development, and providing training and education. There are 4IR development centres and innovation parks planned, as well as initiatives to accelerate investment and adoption in businesses. Several training programmes have been proposed, aimed at a wide range of stakeholders, from students to civil servants.

In terms of mitigating labour displacement, there are initiatives to "provide incentives to minimise the risk of job displacements",[23] "enhance formal social protection mechanism for gig workers"[24] and "gradually reduce foreign labour dependency".[25] Micro, small and medium enterprises (MSMEs) are specifically mentioned as recipients of coordinated support and facilitation to accelerate innovation.[26] The only economic security area that is not directly addressed by the outlined initiatives is the reduction of inequalities, but it was acknowledged in the document that 4IR technologies can widen social and economic inequality.[27]

Political Domain

Many of the initiatives fall under the political domain, but most of them (at least nine) focus on the category of government expertise in AI and digital infrastructure. Within that are a number of services targeted at the government sector—such as MyGovCloud to promote cloud computing in the public sector,[28] a 4IR Innovation Accelerator to drive 4IR adoption at all levels of government,[29] and a Government Experience Lab to drive 4IR innovation.[30] The National Digital Identity programme is expected to catalyse more adoption of 4IR technologies at the state level.[31] In terms of geopolitical strategy and global collaboration, there is a WEF Centre for the 4IR planned, "as a hub of global stakeholders' cooperation to facilitate the development of policy frameworks".[32] 4IR development centres are meant to be "industry-led", so there is definitely public-private partnerships outlined.

23 Initiative 9
24 Initiative 10
25 Initiative 4
26 Initiative 26
27 Page 21
28 Initiative 16
29 Initiative 11
30 Initiative 32
31 Initiative 31
32 Initiative 23

For political security, mis/disinformation and the manipulation of the communication environments were not mentioned within the N4IRP, and neither were checks and balances for surveillance or limitations in power.

Social Domain

Under the social domain, the N4IRP pledges to safeguard society against possible harms by "introducing an ethics framework for technological development, deployment and utilisation",[33] "enhancing personal data protection law, regulations and guidelines",[34] and "introducing specific legislation for cybersecurity".[35] These initiatives are relatively limited in scope, as legal protections are only afforded to personal data protection and cybersecurity issues. The proposed ethics framework, which is not legally binding, seems to cover all other potential harms. In terms of sustainability and ecology, the N4IRP does not discuss the environmental footprint of technology and possible mitigation; what is offered is just support provided to businesses to leverage 4IR technologies to solve socio-environmental issues.[36]

Transparency and accountability of AI systems are not specifically mentioned but as most AI ethical frameworks do cover these,[37] presumably Malaysia's would as well. Human rights are not mentioned within the document. In particular, there is no reference to civil and political rights (CPR), or protections against surveillance. However, as much of the N4IRP focuses on delivering economic, social, and cultural rights (ESCR), it can be argued that the policy does aim to address some aspects of human rights.

Digital/ Physical Domain

Within the digital/physical domain of AI security threats, there are two initiatives that address the issue of cybersecurity, focusing on "introducing specific legislation on cybersecurity"[38] and "enhancing the existing cybersecurity framework by incorporating safeguard measures for the implementation and operationalisation of 4IR across the public sector, with a focus on IoT"[39] (Initiative 25). These do not spell out clearly the aspects of cybersecurity covered, and while "irresponsible use and manipulation of technology" was mentioned a few times in the document as a catch-all phrase for cyber threats, no further elaboration was given. Therefore, potential action could include or exclude any of the digital/physical security risk areas which AI systems can pose a threat to.

33 Initiative 20

34 Initiative 22

35 Initiative 21

36 Initiative 29

37 Anna Jobin, Marcello Ienca, and Effy Vayena, "*The Global Landscape of AI Ethics Guidelines*", Nature Machine Intelligence 1, no. 9 (September 2019): 389–99, https://doi.org/10.1038/s42256-019-0088-2.

38 Initiative 21

39 Initiative 25

This author takes the discretion to decide that the first three areas within the domain, i.e., reliable and value-aligned systems, systems robust against attacks, and protections against malicious use of AI, could be addressed as part of the cybersecurity initiatives and ethical framework as mentioned before, and therefore can be categorized as "possible inclusions". As for the secure convergence of AI and other technologies and AI in warfare, as they are more specific, the assumption is that they are not addressed at this moment.

Discussion

Within this section, we will discuss the assumptions behind the N4IRP and resulting implications on priorities and implementation of Malaysia's technology governance and AI security mitigation.

The Assumptions

The N4IRP's policy wording and slated initiatives point towards a few underlying assumptions. Firstly, even though risks are mentioned, most of the policy strongly suggests that outcomes of 4IR and its associated technologies, including AI, are largely beneficial. For example, stakeholder groups such as businesses, society and government are addressed within the policy as "beneficiary groups" (see Figure 1). Technology is celebrated as progress, its benefits necessarily outweighing the risks. For the most part, the N4IRP reads fairly typically as a policy document, with the language of visions, missions, strategies, and indicators. However, there is a moment in the text where it breaks character and imagines a techno-utopian scenario:

> "Let us take the agriculture sector as an example of the fusion of technologies. A 4IR-ready farmer will oversee a fleet of sensors and robots, and grow tailor-made crops packed with nutrition. The fresh produce will be purchased by consumers from the comfort of their own homes, enabled by the internet and peer-to-peer business models platform. Instead of in-person collection, autonomous vehicles will transport the goods without the need for human travel. Though this scenario may still be years away for some parts of the world, in many places, this is already commonplace." (p.20)

The sense is that Malaysia needs to be heading towards the above scenario, or risk being left behind. Following that, the second apparent premise of the N4IRP is that 4IR is "an inevitable wave of change"[40] that countries will have to adapt to, with urgency. Success will bring about economic growth, competitive advantage, efficiency, and convenience; failure will result in the country losing the developmental race. In order to ride the wave, Malaysia has no choice but to invest heavily in its 4IR ecosystem in the short- and mid-term.

This brings us to the third underlying assumption of the policy: that, with sufficient resources rapidly invested into infrastructure and capacity-building, Malaysia would catch up with countries that are ahead in the technology race, and reap the fruit of its investments. However, this assumption downplays the overwhelming advantage held by other countries in success factors such as talent and innovation ecosystems in the United States, or oceans of data available in China to train and refine its AI models.

Lastly, as is typical in many developmental projects, economic growth is the main indicator and direction, with an implicit orientation towards trickle-down economics. While the rhetorics have shifted towards sustainable development, the majority of the action plans in the N4IRP focus on economic security, with a business-friendly, business-as-usual approach.

The Implications

The assumptions behind the N4IRP bring a set of implications to technology governance and security risk management. Firstly, technology is viewed from the lens of being a solution rather than a potential problem, and therefore most AI security measures within the policy fall within the bucket of mitigating the risk of being left behind, instead of risks connected to safety and unintended consequences, or abuse with malicious intent. While solutions are touted for sustainable development in rhetoric, the main focus remains to be economic competitiveness. 4IR technologies are not scrutinized for the social and environmental problems that they may bring; instead, great faith is placed on technological innovation which may not address systemic and structural causes to the problems.

Secondly, there is a limited approach towards regulation, with an emphasis on speed instead of safeguards. In the N4IRP, regulatory frameworks were mentioned but specifically within the areas of personal data protection and cybersecurity issues, but there was no mention of legislation in areas such as product safety, protection from AI discrimination and bias, algorithmic accountability and transparency in 4IR technology use, just to name a few areas. Throughout the policy, an "anticipatory and agile regulatory approach" was advocated, elaborated within Initiative 19 as regulations to "meet the needs of the digital economy businesses".[41] The proposed

40 Subsection within Chapter One, p.20, N4IRP

41 N4IRP, Page 52

ethics framework seems to be the proposed safety net to address safeguards, but it is not legally binding.

> "4IR technologies are not scrutinized for the social and environmental problems that they may bring; instead, great faith is placed on technological innovation which may not address systemic and structural causes to the problems."

Thirdly, the positioning of 4IR as a key economic enabler to Malaysia's development has certain implications in the governance structure and implementation of the N4IRP. Spearheaded by the Economic Planning Unit which is central to Malaysia's development planning, 4IR and related technologies are elevated into high priority to be mainstreamed across the public sector and civil service. While the lead ministry on digital technologies is the Ministry of Communications and Multimedia (KKMM) and the National Policy on Industry 4.0 (Industry4WRD) focusing on the manufacturing sector is overseen by the Ministry of International Trade and Industry (MITI), these are now consolidated under the National Digital Economy and 4IR Council, led by the Prime Minister.

With six clusters (digital talent, digital infrastructure and data, emerging technology, economy, society, and government) chaired by line ministers and the chief secretary to the government, and relevant ministries slated under the individual clusters, the governance structure embeds a higher level of cross-ministerial and interdisciplinary coordination. Although some have commented that the bureaucracy of the Council may stifle innovation [42] and power dynamics within the Council are yet unclear, it can be argued that some level of friction and feedback loops from relevant ministries and agencies may be beneficial to bring in more holistic considerations and safeguards.

How may this look like in practice? While the N4IRP does not assign lead agencies to policy actions, its sister policy the Digital Economy Blueprint does go to that level of granularity. The MDEB provides an indication on how policy initiatives can be cascaded to ministries that have the mandate and the experience to handle challenges that arise from the digital economy, such as assigning the Malaysia Competition Commission (MyCC) to streamline competition policies and laws, MITI to incorporate digital economy elements into international trade arrangements and negotiations, and the Ministry of Finance to come up with a digital tax framework. These ministries are not traditionally involved in digitalization or technology, but are important for integrating the digital into policy and regulatory frameworks.

42 Siew Yean Tham, "*Malaysia's Digital Economy Blueprint: More Is Not Better*", FULCRUM, 2 March 2021, https://fulcrum.sg/malaysias-digital-economy-blueprint-more-is-not-better/.

Conclusion

Malaysia's N4IRP follows a familiar playbook of investing in economic and human resources to catch up in the technological and developmental race. The policy plan lays out a ten-year plan of "enhancing 4IR awareness and adoption" (two years, in Phase One), "driving transformation and inclusivity" (three years, Phase Two), and "achieving balanced, responsible and sustainable growth by leveraging 4IR technologies" (five years, Phase Three). Mapped against Newman's AI Security Map, misuse and abuse of AI technologies do not weigh heavily in this trajectory, and much remains unsaid within the policy about safeguards, regulatory or otherwise.

AI security risks aside, a techno-utopian vision of Malaysia's future seems simplistic and divorced from realities on the ground. While supporting 4IR technologies is high up in its priorities, there are many local and global challenges that compete for attention and resources within the country. In August 2021, a month after the launch of the N4IRP, the then Prime Minister Muhyiddin Yassin had to resign and dissolve his cabinet following months of political instability. This was the third change of administration in Malaysia within the span of a little more than three years.[43] As such, Muhyiddin Yassin who provided the foreword in the N4IRP is no longer in power. Indeed, in the recent years Malaysia has been fraught with uncertainties including drastic disruptions by the COVID-19 pandemic and the resulting global economic downturn; it is also vulnerable to the climate and ecological crisis which requires much resources for mitigation[44] and adaptation.[45] The N4IRP which advocates a "whole of nation" approach does not mention how the above conditions faced by the public and private sectors in Malaysia, or indeed, the population in general, may hamper the country's abilities to invest in, coordinate on, and benefit from the Fourth Industrial Revolution.

As the N4IRP has just been announced, there is still much room to refine the country's 4IR pathway to focus on resilience rather than rapid adoption. While policy initiatives do focus on narrow economic gains, the interagency governance structure to implement and monitor the N4IRP has the potential to provide the bridging mechanism and expertise across domains to ensure adequate safeguards, so that the pursuit of technology for development does not come at the cost of sustainable development itself.

43 In 2018, the 14th General Election of Malaysia saw an unprecedented defeat of the ruling coalition Barisan Nasional which had governed the country from its independence in 1957, and the regime changed hands. The opposition coalition, Pakatan Harapan, came into power, only to be overthrown two years later in 2020 because some members of parliament changed their party allegiance. The new Prime Minister Muhyiddin Yassin governed for 17 months, during which a state of emergency was announced, suspending parliament and all elections due to the worsening COVID-19 pandemic. The political instability continued towards the end of the emergency in August 2021, when Muhyiddin Yassin resigned after losing majority support of the MPs, paving the way for a new cabinet by current Prime Minister Ismail Sabri Yaakob.

44 The country's plan to be carbon neutral earliest by 2050 was announced during the tabling of the 12th Malaysia Plan in September 2021.

45 Reuters, "*Malaysia to Spend $335 Million for Flood Relief*", Reuters, 29 December 2021, sec. Commodities, https://www.reuters.com/markets/commodities/malaysia-spend-335-million-flood-relief-2021-12-29/.

CHAPTER 03

An Ill-advised Turn: AI Under India's e-Courts Proposal

03

CHAPTER 03

An Ill-advised Turn: AI Under India's e-Courts Proposal

VIDUSHI MARDA

Abstract

In this essay, I analyze the envisioned role and extent of artificial intelligence (AI) applications within the Indian Supreme Court's e-Courts Project. Given the emerging trend of using AI in judiciaries in jurisdictions across the world, this essay studies the current vision of how AI applications are meant to solve problems within the Indian court system and the implications of such deployment. It ends by reflecting on challenges within this vision of e-Courts in India that are emblematic of the current national approach to AI design, development, deployment and governance.

Digital Futures Lab | Konrad-Adenauer-Stiftung

Introduction

Artificial Intelligence (AI) has gained significant momentum as a policy priority in India over the last five years, from the Ministry of Commerce and Industry declaring its intent to use AI to transform India's economy in 2017,[1] to NITI Aayog (a government-run think-tank) publishing a National Strategy on AI in 2018,[2] to the National Automated Face Recognition System (AFRS) being announced in 2019.[3] AI applications like the AFRS are usually introduced and implemented largely through Ministry-specific policies, often under the umbrella of efficiency, cost effectiveness and modernization.

In April 2021, the Supreme Court's e-Committee published a "Draft Vision Document on Phase-III of the e-Courts Project" ("Draft"), envisioning an accessible, efficient and equitable judicial system for all individuals that are part of the delivery of justice in India.[4] Approved in 2007, the e-Courts project was divided into three phases. Phase I ran from 2007 to 2015 and involved the provision of laptops to judicial officers, the development of the Case Information System software (CIS) being made available for deployment, the training of judicial officers and court staff on how to use newly installed softwares, and the national e-courts portal also became operational during this time.[5] Phase II was sanctioned by the Government in 2015 and saw further changes including additional hardware being added to each courtroom, courts being connected to jails via video conferencing, and the launching of the e-Courts services mobile app.[6]

The Draft published in April 2021 conceptualizes Phase III of the project. It proposes an "ecosystem approach" which prioritizes scale, speed and sustainability. The Draft notes, "Given the large, diverse and constantly evolving needs of different users and the constant evolution of technology, administration of justice must not just remain as a sovereign function, but evolve as a service: to mitigate, contain and resolve disputes by the courts and a range of public, private and citizen sector actors."[7] Explaining

1 2017 Artificial Intelligence Task Force, https://www.aitf.org.in/.

2 Niti Aayog. "*National Strategy for Ai: Discussion Paper,*" June 2018, https://smartnet.niua.org/sites/default/files/resources/nationalstrategy-for-ai-discussion-paper.pdf. Also see: "*Responsible AI for All*", Niti Aayog, February 2021, http://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf.

3 Vidushi Marda, "*Facial recognition is an invasive and inefficient tool,*" The Hindu, July 22, 2019, https://www.thehindu.com/opinion/op-ed/facial-recognition-is-an-invasive-and-inefficient-tool/article28629051.ece

4 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf.

5 "*Brief on eCourts Project*", Department of Justice, https://doj.gov.in/sites/default/files/Brief-on-eCourts-Project-(Phase-I-%26-Phase-II)-30.09.2015.pdf.

6 "*eCourts Project, Phase II, Objectives Accomplishment Report, E-Committee Supreme Court of India*", https://ecourts.gov.in/ecourts_home/static/manuals/Objective%20Accomplishment%20Report-2019.pdf; "*eCourtrs Mission Mode Project*", E-Committee Supreme Court of India, https://ecommitteesci.gov.in/project/brief-overview-of-e-courts-project/.

7 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 5.

how this would happen, the Committee notes, "Rather than focus on developing all the solutions itself, the judiciary can curate the right environment and infrastructure for solutions to emerge rapidly from the ecosystem of public and private actors."[8] The "right environment", as per the Draft includes inter alia, designing technology and processes for ease and access of all actors in the ecosystem, namely, litigants, lawyers, registry and civil society, and also designing a system "that enables different parts of the justice delivery system (legal aid authorities, prisons, police etc) to collaborate and provide seamless delivery of justice to citizens by reducing touchpoints."9

The key building blocks of the third phase will enable this turn to an "ecosystem approach" by simplifying procedures, creating foundational digital infrastructure, and setting up Technology Offices at High Courts that will "support the configuration and adoption of the Digital Infrastructure, develop new services, and address grievances".[10] The practical applications that form part of the goals of this phase span a broad range: from ensuring reliable connectivity, to developing a free case law repository to "making machines readable and secure",[11] to enabling E-Filing, transcriptions, open online hearings, remote digital assistance, etc.

This phase of the e-Courts project is fundamentally different from the preceding two. Firstly, Phases I and II focused on making existing processes more efficient by introducing digitization, hardware and software into processes that were hitherto completely offline, whereas Phase III focuses on a fundamental shift, to go beyond the simply digitizing processes. The vision document states, "Given that most judicial processes and procedures evolved in the pre-digital age, it is critical to examine whether such processes continue to remain relevant in a digital age or can be simplified and transformed to better serve justice."[12] Second, the first two phases involved training staff within the judiciary and providing technical infrastructure, among others. The third phase opens up the judiciary to private, public and citizen actors, and focuses on digital technologies that transform the judicial system. Perhaps most importantly, this phase envisions the administration of justice to evolve from a sovereign function to a service.

8 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 20.

9 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 19.

10 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 7.

11 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 39

12 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 28.

In this piece, I show how the Draft gives into tendencies of technological solutionism, i.e. ushering in technologies as a panacea to complex societal problems.[13] The Draft aims for the judicial system to become one that is "natively digital". It contemplates a number of problems that broadly converge around the impact of legacy systems and processes on access to justice and existing pendency in courts.[14] As I will show in the pages that follow, digital technologies are viewed as a solution to these problems due to the classic promise of seamless, efficient, and scalable information analysis and organization. While this seems like an exciting pivot in theory, the sobering reality of technical infrastructure, particularly AI applications discussed in this paper, is that the design, development and deployment of these systems is teeming with complexities surrounding the nature of these systems, their suitability to solve problems that come up in the process of administering justice and the societal impact of use. In practice, the adoption of technology should ideally be iterative, deliberate and only contemplated in tandem with accountability mechanisms, human oversight, and ongoing evaluation.

In its current form, the Draft brings to the fore multiple concerns vis-a-vis access to justice, the doctrine of separation of powers, and the nature and role of the judiciary itself. While extensive comments have been provided by experts on the entirety of the Draft, I will focus on AI-related applications and their implications alone.[15]

India is by no means the only country contemplating the use of AI in the judicial system, although proposed use cases vary between countries. In the US, automated pre-trial risk assessments have been reported for years.[16] In 2019, reports of Estonia's government commissioning a "robot judge" that can investigate small claims disputes emerged in the news.[17] Since March 2020, COVID-19 restrictions have accelerated efforts to digitize courts, with courts in Nigeria and Singapore

13 Natasha Dow Schull, "*The Folly of Technological Solutionism: An interview with Evgeny Morozov*", Public Books, September 9, 2013, https://www.publicbooks.org/the-folly-of-technological-solutionism-an-interview-with-evgeny-morozov/; Vidushi Marda, "*Papering over the cracks: On "Privacy v. Health*", in Taylor, L.; Sharma, G; Martin, A.K.; Jameson, S.M. (eds) Data Justice and COVID-19: Global Perspectives. London: Meatspace Press, 2020.

14 Vidushi Marda, "eCourts in India: Questions facing the Indian Supreme Court's new e-Committee - an interview with Akhil Bhardwaj", AI Now Institute, January 4 2022, https://medium.com/@AINowInstitute/ecourts-in-india-questions-facing-the-indian-supreme-courts-new-e-committee-ef25ec53224a.

15 Internet Freedom Foundation, "*Comments on the Draft Digital Courts Vision and Roadmap Document for Phase III of the e-Courts Project*" , May 13, 2021, https://drive.google.com/file/d/1FXHIwmkVRCxo7PZgCIIHGJeBfp84WX5a/view; Article 21 Trust and others, "*Response to the Draft Vision Document on Phase III of the eCourts Project*", May 31, 2021, https://drive.google.com/file/d/1re-RysqdVtwIVKXtu8ZfHSOESQb-h7V_/view; Siddharth Peter DeSouza, Vasha Aithala and Srishti John, "*The Supreme Court of India's Vision for e-Courts: The Need to Retain Justice as a Public Service*", The Hindu Centre Policy Watch No. 14, July 10, 2021, https://www.thehinducentre.com/publications/policy-watch/article34779031.ece.

16 Julia Angwin et al, "*Machine Bias*", ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Richard M. Re & Alicia Solow-Niederman, "*Developing Artificially Intelligent Justice*", 22 Stanford Technology Law Review 242 (2019), https://www-cdn.law.stanford.edu/wp-content/uploads/2019/08/Re-Solow-Niederman_20190808.pdf.

17 Eric Niiler, "*Can AI be a fair judge in Court? Estonia thinks so*", WIRED, March 25, 2019, https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/.

reporting one instance each of courts sentencing a person to death over Zoom.[18] China's Smart Courts ecosystem has steadily grown over the last few years, with the smart court system being folded into China's National Strategy for the Informatization Development in 2016 which standardizes and guides national informatization development in the country.[19]

The use case of AI in courts is important to examine given the profound impact it will have on access to justice and structural inequality. Secondly, examining the language and intentions laid down in the draft can tell us about foundational assumptions at play and can help in understanding how technocratic tendencies manifest across the spectrum of AI policies in India.

> "In practice, the adoption of technology should ideally be iterative, deliberate and only contemplated in tandem with accountability mechanisms, human oversight, and ongoing evaluation."

This essay will proceed as follows. In the next section, I will discuss the intended role for AI systems within the Draft, focusing on three aspects: the emphasis on "intelligent scheduling", the intended scale adoption of SUVAS (Supreme Court Vidhik Anuvaad Software) to translate judicial documents, and the creation of a digital infrastructure that enables an Interoperable Criminal Justice System, and analyze issues arising from these proposed use cases. Following this, I zoom out to analyze these intended developments against the backdrop of AI governance in India.

18 Rebecca Ratcliffe, "*Singapore sentences man to death via Zoom call*", The Guardian, May 20, 2020, https://www.theguardian.com/world/2020/may/20/singapore-sentences-man-to-death-via-zoom-call; Kechi Nomu, "*Death decreed over Zoom*", Rest of World, September 14, 2020, https://restofworld.org/2020/death-decreed-over-zoom/.

19 Shazeda Ahmed, "*In the Shadow of the 'Smart Court' - Examining China's Applications of Courtroom AI*", Stanford HAI, November 12, 2020, https://www.youtube.com/watch?v=t8KONYsWn6k&t=2239; Claire Cousineau, "*Smart Courts and the push for technological innovation in China's judicial system*", Center for Strategic and International Studies, April 15, 2021, https://www.csis.org/blogs/new-perspectives-asia/smart-courts-and-push-technological-innovation-chinas-judicial-system; Changqing Shi, tania Sourdin & Bin Li, "*The Smart Court - a new pathway to justice in China?*", International Journal for Court Administration, 12(1), DOI: http://doi.org/10.36745/ijca.367; State Council General Office, "*Outline of the National Informatization Development Stategy*", China Copyright and Media, July 27 2016, https://chinacopyrightandmedia.wordpress.com/2016/07/27/outline-of-the-national-informatization-development-strategy/.

India's Vision for e-Court: Where Does AI Come In, and Why?

Efficiency, equity, access and inclusion are values that find mention throughout the document, referred to as the "founding vision" of Phase III of the e-Courts project. A number of applications proposed within the Draft threaten to weaken this founding vision, firstly because the Draft does not engage with the nature of AI systems even as it proposes their use, and secondly, because it does not to engage with everyday realities of the judiciary beyond just gaining access to courts, which undermine equity and inclusion in the judicial system.

Intelligent Scheduling

The Draft refers to "intelligent scheduling" multiple times, emphasizing that this Phase III effort towards creating a digital infrastructure will "allow for greater predictability and optimization of capacity of judges and lawyers" by enabling "data-based decision making for judges and registries when scheduling or prioritizing cases". There are hints that intelligent scheduling is only the first of many such applications, as the Draft also heralds "a future of macro data-driven decision making enabling targeted interventions and resource allocation both on the judicial and administrative side."[20]

The purported benefits of intelligent scheduling abound throughout the document. It is proposed as a transformative technology that can reduce the cognitive burden on judges by managing legal aid services by digitally allotting cases and help evolve services like queue management for lawyers.

However, using AI to help with these problems reveals crucial gaps in the Committee's understanding of the nature and limitations of the technology. Discussing legal aid services in India, Upendra Baxi stated: "Solicitude for the poor calls for hard headed appraisal of social realities and sober formulation of feasible policies", an approach that the Draft abandons in favour of "efficiency" introduced by AI technologies. [21] Further, AI systems lack an understanding of context and cultural

20 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 5.

21 Upendra Baxi, "*Legal Assistance to the Poor: A Critique of the Expert Committee Report*", Economic and Political Weekly, Vol. 10, No. 27 (Jul. 5, 1975), pp. 1005-1013 (9 pages) , https://www.jstor.org/stable/4537237?seq=1.

context.[22] In the Indian judicial system, it is not just judges or lawyers that bear the burden of poorly scheduled cases, but litigants themselves.[23] Considerations like proximity to courts, the social and economic realities of litigants, the subjective and personal barriers to approaching courts, etc., need to be factored in at the time of assessing the urgency and scheduling of a case, which necessarily requires a nuanced understanding of the need and stakes at play. The choice of how cases should be scheduled, what constitutes urgent matters and how urgency between multiple matters is calibrated are deeply subjective, although the Draft does not acknowledge or discuss this. It notes, "intelligent scheduling can generate data to identify cases that need to be prioritised, and generate data and act as a capability to support digital listing and other services."[24] The process of improving scheduling within courts includes understanding specific needs of litigants, the merits of specific cases, and creating a hospitable environment for parties to have access to justice—a technical solution like intelligent scheduling may help when used in a very limited capacity, but the idea that it can identify cases to be prioritized is ill-informed, particularly in the absence of any discussion on the qualitative considerations or underlying causes of problems at hand.[25] Relegating them (even partly for the purposes of recommendation) to the realm of machine-driven decision-making ignores the complexities involved in ascertaining the bearing of each case on the lives and interests of litigants, alluding to it being a much more straightforward task than it really is in practice.

SUVAS

The Supreme Court Vidhik Anuvaad Software, abbreviated to "SUVAS", was first presented to the President of India on Constitution Day, November 26, 2019.[26] Intended to translate English judicial documents, judgments and orders into nine vernacular languages, it is proposed as a key goal under the e-Courts vision document.

Machine translation requires that the machine translates the meaning from one language to another fluently, which has proved to be a complex task even with

22 Thant Sin, "*Facebook bans racist word 'Kalar' in Myanmar, triggers censorship*", Business Standard, June 3, 2017, https://www.business-standard.com/article/international/facebook-bans-racist-word-kalar-in-myanmar-triggers-censorship-117060300423_1.html; Vidushi Marda, "Regulating social media content: Why AI alone cannot solve the problem", ARTICLE 19, July 11, 2018, https://www.article19.org/resources/regulating-social-media-content-why-ai-alone-cannot-solve-the-problem/.

23 Nick Robinson, "*Hard to reach*", Frontline, February 12, 2010, https://frontline.thehindu.com/other/article30179161.ece.

24 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 39-40.

25 Pratik Datta and Suyash Rai, "*How to start resolving the Indian Judiciary's long-running case backlog*", Carnegie Endowment for International Peace, September 9, 2021, https://carnegieendowment.org/2021/09/09/how-to-start-resolving-indian-judiciary-s-long-running-case-backlog-pub-85296; also see Article 21 Trust's response, supra note 5.

26 Supreme Court of India, Press Release, November 25, 2019, https://main.sci.gov.in/pdf/Press/press%20release%20for%20law%20day%20celebratoin.pdf.

cutting-edge machine translation systems. For instance, in April 2020, Google's state-of-the-art translation software translated an English sentence referring to a court enjoining violence, to a sentence in Kannada that stated that the court ordered violence.[27] Researchers from Carnegie Mellon caution against the use of machine translation technologies without careful thought and planning, stating: "The recent raft of high-profile gaffes involving neural machine translation technology has brought to light the unreliability and brittleness of this fledgling technology... we present this cautionary tale where we shed light on the specifics of the risks surrounding cavalier deployment of this technology by exploring two specific failings".[28] There is no mention of the limitations or fallibility of these technologies, even though machine translation has had modest success in being reliable enough for real-world application thus far.

Even so, the Draft entails SUVAS being used to "enable integration of data across languages from prisons and police stations" and "digitally assisted language translation must form the basis for redesigning some of the most effort and time-intensive administrative processes." State-of-the-art technology offered by Google has been found to be unfit for medical translations, given the mistakes and impact of people's lives, and judicial systems should think twice before automating these processes.[29] According to information furnished from records of unstarred questions in the Rajya Sabha, it appears that SUVAS, as of February 2020, was being tested, trained and refined in 18 High Courts, although more information about results of initial tests, accuracy and reliability are not known at the time of writing this essay.[30]

Machine translation is clunky and unreliable for a number of reasons. First, these systems do not understand aspects of common sense that is usually applied when we translate between languages, meaning that there is a lot lost in such translation, and sometimes, translation can also result in bad outcomes, as evidenced by the Google example above. Second, machine translation comes with low levels of accuracy, particularly when word-to-word translations are required given the lack of understanding of context and nuance.[31] Because of this, even if machine translation is quick and "efficient", it still requires a human eye to go through translations with a fine-toothed comb, resulting in duplication of efforts. It is pertinent to note that the Draft does not contemplate human oversight, a costly omission that merits close

27 Paresh Dave, "*Google AI Translate botches legal terms 'enjoin' 'garnish' - research*", Reuters, April 19, 2021, https://www.reuters.com/technology/google-translation-ai-botches-legal-terms-enjoin-garnish-research-2021-04-19/.

28 Vinay Prabhi, "*Google translate bias investigations*", YouTube, https://www.youtube.com/watch?v=eKRpiMBIu40; Dr. Vinay Prabhu et al, "*Did they direct the violence or admonish it? A cautionary tale on contronomy, androcentrism and back-translation foibles*", AfricaNLP 2021, https://openreview.net/pdf?id=hUzjN3Sjrc.

29 Nicole Wetsman, "*Google Translate still isn't good enough for medical translations*", The Verge, March 9, 2021,https://www.theverge.com/2021/3/9/22319225/google-translate-medical-instructions-unreliable.

30 Department of Justice, Rajya Sabha, Unstarred question number 587: https://doj.gov.in/sites/default/files/RJ-Hindi_16.pdf.

31 Doughlas Hofstadter, "*The Shallowness of Google Translate*", The Atlantic, January 30, 2018, https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/

inspection, especially as real-world examples demonstrate the need for oversight and careful review. The focus on "seamless" access to translated documents and "reducing touchpoints" through the document are also a cause of major concern. Particularly as researchers have pointed out that in the Indian context, machine translation is still facing significant challenges, and are only in the process of being better understood and ironed out.[32]

Algorithmic Transparency

The Draft cursorily discusses the issue of algorithmic transparency. In the context of core values guiding this phase, it states it is important that the "digitisation efforts should ensure that constitutional and legal rights accorded to individuals, of dignity to life, liberty, equality, freedom and fraternity are guarded and secured. They should enhance the trust and ability of the legal system to secure the rights of individuals. This demands that the process of digitisation is consultative by design, inviting inputs from all. Equally that digitisation processes advance trust by enabling and leveraging ecosystem capability to serve justice". The Draft contemplates that adopting open-source software and algorithmic transparency will enhance trust.[33]

This is a perplexing stance to take in context of wider AI applications discussed in the Draft for two reasons. First, fundamental rights considerations are peripheral (at best) to proposals for re-engineering processes or creating a new digital infrastructure throughout the Draft, and second, the inherent opacity and inscrutability of algorithmic systems is a rich field of study, and yet the tension this poses is wholly ignored in the Draft. Algorithmic transparency, in particular, is a can of worms as some scholars argue that transparency as an end in and of itself means very little—having access to source code, or even training data does little to establish trust or demonstrate predictability.[34] The true test of algorithmic systems that are able to establish trust are those that lend themselves to accountability mechanisms, ongoing scrutiny, testing, and audits—none of which are even contemplated in the Draft.[35]

Interoperable Criminal Justice System

One of the key goals of Phase III of the e-Courts project is to prioritize core digital platforms, and includes operationalizing and scaling the Interoperable Criminal

32 Raj Nath Patel, Prakash B Pimpale and M. Sasikumar, "*Machine Translation in Indian Languages: Challenges and Resolution*", Journal of Intelligent Systems, https://doi.org/10.1515/jisys-2018-0014, https://www.degruyter.com/document/doi/10.1515/jisys-2018-0014/html#j_jisys-2018-0014_s_005_w2aab3b7b6b1b6b1ab1b4Aa.

33 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 24.

34 Vidushi Marda, "*Machine Learning and Transparency: A Scoping Exercise*", November 22, 2017, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3236837.

35 Inioluwa Deborah Raji et al, "*Closing the AI Accountability Gap Defining an End-to-End framework for Internal Algorithmic Auditing*", In Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3351095. 3372873.

Justice System (ICJS). The ICJS was launched in 2019 with a view to promote seamless data exchange and analytics between various parts of the criminal justice system. Housed under the crime and criminal tracking network system (CCTNS) project at the Ministry of Home Affairs, the ICJS integrates data from police, forensics, e-prisons and e-Courts.[36] By virtue of its structure, it implicates various parts of India's surveillance landscape, given that these pillars are also linked to databases like the National Database on Sexual Offenders (NDSO) and the proposed National Automated Face Recognition System (AFRS), among many others.

It is important to bear in mind that India does not have a data protection law, and the current Data Protection Bill contemplates wide exemptions for State actors.[37] This means that data sharing, storage and retention within the ICJS—and its various building blocks—will operate with little oversight. Let us consider a single building block within the ICJS—the AFRS. The AFRS is contemplated to be a centralized database used for criminal identification and verification, based on CCTNS data, along with pictures from police raids, newspapers, among others, to identify and verify pictures of people. In the absence of data protection safeguards, the AFRS may be used for purposes that go beyond its intended scope. For instance, in December 2019, Delhi Police used facial recognition to monitor peaceful protests, and claimed that the legal basis for this arose from a Delhi High Court judgment instructing the government to use facial recognition technology to find missing children. In a blatant case of function creep, it was seamlessly applied to monitor people as they exercised constitutional rights. This is additionally worrying given demonstrated shortcomings of facial recognition systems and the inherent unconstitutionality of its aims and structure.[38] It is being proposed in the absence of a valid legal basis and, more recently, its use for surveillance at public congregations has come under scrutiny for being unconstitutional.[39]

36 Ministry of Home Affairs, "*She-Raksha*", April 1 - October 15 2019, https://www.mha.gov.in/sites/default/files/WSDivision_SheRakshaVol2_08112019pdf.pdf.

37 Sobhana K. Nair, "*UIDAI wants exemption from Data Protection Bill*", The Hindu, October 29, 2021, https://www.thehindu.com/news/national/uidai-wants-exemption-from-data-protection-bill/article37238680.ece

38 Vidushi Marda, "*Every Move You Make*", India Today, November 29, 2019, https://www.indiatoday.in/magazine/up-front/story/20191209-every-move-you-make-1623400-2019-11-29; Vrinda Bhandari, "*Facial recognition: Why we should all worry about the use of big tech for law enforcement*", The Future of Democracy in the Shadow of Big and Emerging Tech (CCG, NLU Delhi/FNF, 2021), https://ccgdelhi.s3.ap-south-1.amazonaws.com/uploads/the-future-of-democracy-in-the-shadow-of-big-and-emerging-tech-ccg-248.pdf; Smriti Parsheera, "*Adoption and regulation of facial recognition technologies in India: Why and why not?*," Data Governance Network, April 16, 2020, https://datagovernance.org/report/adoption-and-regulation-of-facial-recognition-technologies-in-india.

39 S. Q. Masood v. State of Telangana, Memorandum of Writ Petition, 191/2021, High Court for the State of Telangana in Hyderabad, https://drive.google.com/file/d/1cQdzjT8mW0VRwtJh1shW0RKWQ4k4J1sM/view; Vidushi Marda, "*Every Move You Make*", India Today, November 29, 2019, https://www.indiatoday.in/magazine/up-front/story/20191209-every-move-you-make-1623400-2019-11-29; ; Internet Freedom Foundation, "N*otice to cease use of facial recognition technologies by the Delhi Police as it is an illegal act of mass surveillance*", Sent to the Ministry of Home Affairs, December 28 2019, https://drive.google.com/file/d/1-GA-LIcVInm0Ln4nuA_E_gLBMr6zeWzn/view.

The ICJS intends to integrate data from courts, prisons, police and forensics. This proposed blurring of lines between the executive and the judiciary also brings into question the separation of powers laid down in the Indian Constitution. As the Article 21 Trust response to the Supreme Court e-Committee notes, "the judiciary, particularly the constitutional courts, not only has to be in reality independent of the executive, but also seen to be so. Any measure that even remotely suggests or portrays that the judiciary and the executive are working in tandem or in unison ought to be avoided and the separation therefore has to both be real & apparent."[40]

Further, police data is imbued with evidence of historical discrimination particularly along the lines of caste, low income groups and women.[41] The process through which data collection and creation happen within policing institutions also exhibits biases of various kinds, from the nature of data collected, to the characterization afforded to data about reported crimes and incidents.[42] Prior to the ICJS, as lawyers Nikita Sowane, Srujana Bej and Ameya Bokil observe, "Given that the system cannot yet 'flow freely', police have to officially submit these records before the courts, giving the accused the right to challenge the correctness of the record. But an interoperable system creates potential for this information to be used to the detriment of accused persons without their knowledge." [43]

Even judging from a handful of components that make up the ICJS, the assumption that it passes legal muster is a flawed one to begin with—and centring these systems as a part of the e-Courts project only exacerbates the problems posed by what this interconnected future means for individual rights and rule of law.

Analysis

Painting With An (Uncritical) Broad Brush

The Draft suggests that technology presents appropriate solutions to the problems of access to justice, procedural inefficiencies within courtrooms, increasing burden on courts, nuances of decisions taken at the stages even preceding the commencement

40 Article 21 Trust and others, "*Response to the Draft Vision Document on Phase III of the eCourts Project*", May 31, 2021, https://drive.google.com/file/d/1re-RysqdVtwIVKXtu8ZfHSOESQb-h7V_/view

41 Ameya Bokil et al, "*Settled Habits, New Tricks: Casteist Policing Meets Big Tech in India*", Long Reads, May 2021, https://longreads.tni.org/stateofpower/settled-habits-new-tricks-casteist-policing-meets-big-tech-in-india; Maja Daruwala, "*Fair and Unbiased Policing Still a Far Cry in India*", The Wire, June 4, 2018, https://thewire.in/society/fair-and-unbiased-policing-still-a-far-cry-in-india.

42 Vidushi Marda and Shivangi Narayan, Data in New Delhi's Predictive Policing System, FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 2020, DOI 10.1145/3351095.3372865, https://www.vidushimarda.com/storage/app/media/uploaded-files/fat2020-final586.pdf

43 Nikita Sonavane, Srujana Bej and Ameya Bokil, "*The dangers of a centralised database for justice system*", The Indian Express, May 28, 2021, https://indianexpress.com/article/opinion/columns/the-dangers-of-a-centralised-database-for-justice-system-7333252/.

of judicial proceedings, and of providing parties with estimates of fair compensation, options for legal recourse, and so on.

It is imperative to note that the Draft paints with a dangerously broad brush, i.e. it fails to reckon with the reality that not all problems can or should be solved by technology, and tends to evade questions of suitability of technical solutions in crucial cases. In other words, it falls into the "Solutionism trap" defined by Selbst et al as: "Failure to recognize the possibility that the best solution to a problem may not involve technology".[44] For example, using algorithmic systems to provide input on what fair compensation looks like in a particular case does not inspire confidence in the fairness of that process, even in the case of compensation norms "conclusively settled by statute or case law".[45] Given the context-specific facts of the case, litigants, etc, even seemingly straightforward cases require the application of judicial minds—it is not appropriate to relegate this to the realm of machine learning alone. Similarly, the problem of scheduling hearings is not merely a matter of optimizing the capacity of lawyers and judges, and understanding availability of witnesses, case type, etc. An understanding of the struggles of litigants, barriers to participation for them, the urgency of the case based on its merits are all crucial considerations that must go into scheduling. The problem is not solved by technology and may, in fact, be overlooked by an uncritical reliance on technical systems.

This is not to say that technology cannot be useful—but rather to point to the fact that the adoption of technology must be deliberate, bespoke and accompanied by ongoing accountability mechanisms and oversight. Rudimentary safeguards which are necessary but not sufficient, such as human oversight, redressal mechanisms are not discussed in the Draft, even cursorily.

> "It is imperative to note that the Draft paints with a dangerously broad brush, i.e. it fails to reckon with the reality that not all problems can or should be solved by technology, and tends to evade questions of suitability of technical solutions in crucial cases."

44 Solutionism trap in Andrew D. Selbst et al, Fairness and Abstraction in Sociotechnical Systems, FAT* '19: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 2020, DOI https://doi.org/10.1145/3287560.3287598, https://dl.acm.org/doi/pdf/10.1145/3287560.3287598.

45 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 29.

Ambiguity and Empowerment: A Double-Edged Sword

The Draft is also ambiguous in instances where the role of "transformative" technologies is mentioned. At the same time, the unfulfilled promises of such transformative technologies are emphatically (even if uncritically) proclaimed. For instance, in discussing the digital case registry, the Draft notes that it can help with predictable services like e-filing, and also "allow the generation of data to create new parameters for judicial performance evaluation and support intelligent law-making to avoid or contain disputes".[46] The implications of such statements are both vague and unhelpful as it is unclear what these "new parameters" entail, or what is meant by "intelligent law making", or for that matter, which "transformational" technology will be used, procured, tested. This makes it difficult to critically examine what the Draft is proposing. Elsewhere, the Draft is unclear about exactly what kind of technology is intended where. For instance, the Draft states that one of its aims is to "build a 'smart' system, in which registries will have to minimally enter data or scrutinize files owing to foundational capabilities of data connected through leveraged technology".[47] It is unclear whether this refers to algorithmic systems declaring the justiciability of cases as mentioned elsewhere, or if a different kind of "leveraged technology" is being referred to here. The road to clumsy technical deployments is paved with vague promises and lack of oversight, a reality that the Draft clearly does not recognize.

One Fell Swoop: Mischaracterizing the Nature of Technology & Law

Decisions made within the judiciary, from deciding which cases are to be listed, whether a judicial process should be pursued, substantive decisions made in the process of adjudicating cases are all value-laden decisions of equity, fairness and interpretations of the law. In pursuing technology to make these processes efficient, where a system can combine "the vast body of judicial data to foster legal literacy and furnish information on remedies to an aggrieved person at the click of a button" or "determine fair compensation and help avoid or contain disputes", the Draft actively ignores the inherent nature of both technology and the judicial system.[48] In fact, it even proposes using technology at the pre-filing stage to determine a citizen's options for legal recourse, claiming that "well-designed forms with questions and drop-down responses, written in accessible language, can be used to ascertain relevant details such as the cause of action and the value of the suit (if relevant).

46 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 30.

47 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 5.

48 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 5 and 29.

Using appropriate data from the eCourts database, the portal can then provide the citizen with information on the justiciability of the case, the court with jurisdiction over it, and applicable legislation".[49]

These are peculiar issues to relegate to the realm of technology, given that lawyers, judges, litigants often have differing views on such matters. The facts of a case and its bearing on law can be interpreted in a multitude of ways, and the idea that there is a feasible way to outsource such interpretation to a technology with drop-down menus is inherently problematic.

Justice as a Service

The Draft lays down that under the ecosystems approach, "Given the large, diverse and constantly evolving needs of different users and the constant evolution of technology, administration of justice must not just remain as a sovereign function, but evolve as a service: to mitigate, contain and resolve disputes by the courts and a range of public, private and citizen sector actors".[50]

This brings up a host of issues. Firstly, the Draft does not consider how this wider range of actors, including private actors, will be held to account and be subject to oversight and scrutiny. This is particularly important if private actors are expected to have access to multiple databases containing sensitive personal information from police, prisons, and courts (among others). Secondly, justice transcends mere resolution of disputes. As lawyers Siddharth de Souza, Varsha Aithala and Srishti John argue, Courts serve a social function because they provide binding and authoritative decisions that protect rights. Quoting Hazel Genn, de Souza et al. point out, that civil justice is a public rather than a private benefit. Courts of record set precedent, and courts in general are also meant to protect aggrieved parties—to reduce this role to that of a "service" is a gross mischaracterization. Justice is inextricably linked to values such as fairness, accountability, empathy, and rule of law, none of which are adequately considered in the Draft.

This is in stark contrast to the approach of the Draft, which seeks to contain, avoid or mitigate disputes, encouraging individuals to pursue alternate dispute resolutions too. This is consistent with another tendency within the Draft, to sneak in business language while describing ideal changes within the judiciary—at various points, the Draft discusses ways to reduce touch-points within various parts of the judicial system, and terms citizens as "users". It also promotes the idea of "market operators" developing solutions to myriad problems within the judicial system.

49 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 46.

50 "D*igital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 5.

> "The facts of a case and its bearing on law can be interpreted in a multitude of ways, and the idea that there is a feasible way to outsource such interpretation to a technology with drop-down menus is inherently problematic."

Walking Before Running

The Draft acknowledges that even after the first two phases of the e-Courts project, there "remain challenges in ensuring capability, integration of technology and data, and most importantly, adoption". It also acknowledges that basic building blocks of infrastructure like reliable connectivity are lacking in some courts. It even notes that lawyers and litigants themselves may not have the capability to access and use technology. The Draft notes, rather reductively, that this failure to harness technology to its full potential has created a "mind-set barrier against technology services and solutions".[51] Perplexing, because if anything, these realities foresee why technology cannot bring about equity in an unequal society that requires rethinking the status quo of institutions and addressing problems at the root.

The ground realities of how courts function on a day-to-day basis is vastly different from the technocratic vision consistently promoted through the Draft. Much like other policy proposals in India that contemplate predictive policing, facial recognition, AI in education, among others which aim to solve social problems through "modern" technology akin to magic, the Draft seems to fall into the trap of ignoring the very real limitations of technology in favour of unfulfilled potential.

While discussing challenges, the Draft also notes, "there are limited frameworks available for organised feedback resulting in various stakeholders remaining alienated from the [eCourts] system and being passive users." The problem of exclusion and perpetuation of power dynamics that dictate an individual's access to justice within courts is well recognized as a central problem of the Indian judiciary, which requires a careful recalibration of how to organize the justice administration system better.[52] This alienation is not merely a matter of procedural access, but rather of everyday attitudes, assumptions and treatment of marginalized groups. As Constitutional scholar Aparna Chandra states, "Judicial reform measures for improving access to justice tend to focus on overcoming barriers to getting into court,

51 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 18.

52 S. Muralidhar, "*Access to Justice*", Seminar Magazine, https://www.india-seminar.com/2005/545/545%20s.%20muralidhar1.htm.

rather than the treatment meted out within the court system, or on just outcomes from the legal system." She further notes, "Conflating access to justice with access to courts is based on an assumption that courts are necessarily just institutions. This does not take into account the profound sense of alienation felt by marginalised communities, from all state processes, including court processes. Communities that are vulnerable to exploitation and rights violations, whose survival is threatened by socio-economic structures supported by legal injunctions that deny them their basic needs, often see courts as part of the apparatus that serves to keep them disempowered. Marginalised and vulnerable populations are often brought to court as accused persons deserving punishment, rather than coming to court to vindicate their rights." [53]

Zooming Out: Interplay with AI Governance in India

The e-Courts vision is emblematic of a few existing problems at the core of India's AI policy landscape, and distilling them can serve as a prompt for future work aimed at bringing robust safeguards and deliberation into the fold.

Fairness, transparency and accountability

By and large, three fundamental values guide the development and assessment of AI systems: fairness, accountability and transparency.[54] These find mention in NITI Aayog's National Strategy for AI discussed earlier in this essay,[55] and are guiding principles under Indian Constitutional law.[56] The Draft, however, compromises on these in favour of "efficiency".

Fairness encompasses both procedural and substantive fairness, and the over-reliance on technical tools to speed up justice delivery seem to follow the increasingly tainted adage of "move fast and break things". This is particularly dangerous as two levels of fairness are at play— one, at the level of the law and courts, and two, within the technical system itself. Encoding values of fairness into AI systems

53 Aparna Chandra, "*Indian Judiciary and Access to Justice: An Appraisal of Approaches*" , DAKSH India, https://dakshindia.org/state-of-the-indian-judiciary/33_chapter_18.html.

54 While initiatives surrounding these values exist in government, corporate, academic and civil society spaces, a good place to start understanding the vast field of research is the ACM Conference on Fairness, Accountability and Transparency (ACM FaccT): https://facctconference.org.

55 Niti Aayog. "*National Strategy for Ai: Discussion Paper,*" June 2018, https://smartnet.niua.org/sites/default/files/resources/nationalstrategy-for-ai-discussion-paper.pdf.

56 Articles 14 and 16 of the Indian Constitution check against arbitrary State action, laying down safeguards to ensure fairness and equality. Fairness is also assured by the Constitution in the context of procedural fairness. See https://indiankanoon.org/doc/237570/. For a longer discussion on fairness, accountability and transparency as conceived within the realm of machine learning and links to Indian Constitutional Law, see Vidushi Marda, Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making, October 2018. 376 Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. Available from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3240384.

in particular and algorithm systems in general is a complex task, with little to no consensus on what fairness actually means.[57] Multiple definitions of fairness exist, and the challenge of ensuring non-discrimination is an active field of research across stakeholder groups.[58] It isn't (and shouldn't be) a straightforward task. Even so, the Draft remains silent on fairness at both levels.

> "Encoding values of fairness into AI systems inparticular and algorithm systems in general is a complex task, with little to no consensus on whatfairness actually means.[57]"

Similarly, the emphasis on private sector actors providing solutions that will bring about a seamless system runs afoul the principle of transparency. Firstly, because AI systems are inherently opaque and cannot always provide answers for why specific decisions were made.[59] And secondly, because there is little transparency contemplated about how these systems will be built. Will the individuals supposedly benefiting from these systems, i.e. judges, lawyers, clerks and citizens merely be end users, or will they have some level of involvement in how they are ultimately built? Is this neatly pushed into the decision-making realm of private players alone who decide the contours within which systems will function? To what extent can these systems be scrutinized and audited?

Even in the context of live-streaming—the balance between enabling courts to be more open on one hand, and the privacy of judges, litigants and court staff and making hearings more consistently accessible, on the other, is not a tension that the Draft grapples with. Experiences from the United Kingdom, Australia and China suggest that live-streaming may have been necessary during the COVID-19 pandemic, however, the question of whether this virtual turn will be consistently applied across the board, or whether it will be subject to the arbitrary approvals needs to be examined carefully, especially in light of recent decisions by the Central

57 Ben Green. "*The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness*," Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*). 2020, https://scholar.harvard.edu/bgreen/publications/false-promise-risk-assessments-epistemic-reform-and-limits-fairness; Sam Corbett-Davies & Sharad Goel, "*The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*", Computers and Society, arXiv:1808.00023, https://arxiv.org/abs/1808.00023.

58 Arvind Narayanan, "*Tutorial: 21 fairness definitions and their politics*", ACM Conference on Fairness, Accountability and Transparency, March 1, 2018, https://www.youtube.com/watch?v=jIXIuYdnyyk.

59 Cynthia Rudin, "*Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*", Nature Machine Intelligence 1, 206 - 215 (2019), https://www.nature.com/articles/s42256-019-0048-x.

Government.[60] The question of whether live-streaming is always appropriate also needs to be examined carefully as a number of issues have arisen from live-streaming—the existing digital divide in India means that some parties will find it virtually impossible to navigate an online ecosystem, a requirement that shouldn't be imposed on individuals seeking justice. Secondly, live-streaming requires a stable and reliable internet connection and appropriate hardware, which can be a stark contrast from ground realities in Indian courts.[61]

At multiple places in the Draft, the need for redressal mechanisms for technology-related grievances is underscored, as is the need for buy-in from stakeholders. At the same time, the question of how the technologies contemplated in the Draft will be held accountable, remains almost entirely unconsidered. Accountability requires a number of building blocks to be put in place before it can be achieved—transparency of operations and technical systems, standards against which performance will be scrutinized, scrutability of decisions and ongoing audits, real-time assessment of how the system functions and its impact on people and processes, appeal and complaints mechanisms—none of which are adequately considered in the Draft. While "trust" is a key value underlying the building blocks of this phase, it seems to be merely declared instead of demonstrated—trust in systems comes from decisions being explicable, predictable, subject to change, and open to scrutiny, none of which are currently the case.[62]

The current vision proposes "open APIs and also standards, specifications and certifications that can act as guardrails as different actors build solutions across space and time".[63] This does not reflect the centrality of fundamental rights, access to justice and principles of natural justice that are supposed to be inherent to the judicial system.

On Fundamentals

India does not currently have a data protection law, and yet, is home to a burgeoning ecosystem of so-called advanced AI networks from facial recognition to predictive policing to smart judicial applications. As a result, private and State actors develop

60 Prasanth Jha, "Same-sex marriage case in Delhi High Court not of national importance, majority not affected: Central government opposes live-streaming", Bar and Bench, May 16 2022, https://www.barandbench.com/news/same-sex-marriage-case-delhi-high-court-not-national-importance-majority-not-affected-central-government-opposes-live-streaming.

61 Changqing Shi, tania Sourdin & Bin Li, "*The Smart Court - a new pathway to justice in China?*", International Journal for Court Administration, 12(1), DOI: http://doi.org/10.36745/ijca.367.

62 Cynthia Rudin, "*Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*", Nature Machine Intelligence 1, 206 - 215 (2019), https://www.nature.com/articles/s42256-019-0048-x; Andrew D. Selbst and Solon Barocas, "*The Intuitive Appeal of Explainable Machines*", Fordham Law Review (2018). https://ssrn.com/abstract=3126971; Vidushi Marda, "*Machine Learning and Transparency: A Scoping Exercise*", August 31 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3236837.

63 "*Digital Courts Vision and Roadmap: Phase III of the e-Courts Project*", E-Committee Supreme Court of India, https://cdnbbsr.s3waas.gov.in/s388ef51f0bf911e452e8dbb1d807a81ab/uploads/2021/04/2021040344.pdf, at Page 6.

and deploy applications in the absence of regulation or constraints with respect to collecting, storing, processing and sharing personal data. If the Draft Data Protection Bill is any indication of how State actors in particular will be regulated in the future, the answer is bleak—wide exemptions are afforded to the State under the umbrella of vague terms like "sovereignty and integrity of India" and "public order", among others.[64] This demonstrates the general Indian policy approach to put the AI cart before the regulatory horse. This hastiness, however, is only a symptom. A major underlying cause for this is an over-reliance on private actors to suggest solutions and market technologies to the government in the absence of meaningful civil society involvement. In this case, the Draft is several notches above other AI initiatives, given the time and space afforded to comments and consultation.[65] However, the language of the Draft, and assumptions embedded within reveal that the problem is not solved simply by budgeting in windows for civil society comments, and the approach of solutions suggested within are emblematic of the "technology first, society second" mindset pervasive through Indian policy documents on AI.

Network of (Premature) Mutually Reinforcing Frameworks

The legitimacy of AI programmes, from the AFRS to the ICJS and to the e-Courts project, seems to be reinforced by the enmeshed vision independently perpetuated in each of these proposals. They originate and find their legal and financial basis in policy documents, as opposed to legislative proposals. This explains the limited reckoning with the legitimacy, legality, proportionality and necessity of investing in these systems in the first place. This exposes India's AI policy framework to being legitimized not through legal oversight and scrutiny, but rather by investment and unfettered use—the political and financial incentive to roll back technologies after crores have been invested in procuring them declines sharply over time. This points to the urgent need for Indian AI governance to pause and reflect on the inherent nature of AI systems, their limitations and appropriateness in supplanting various State functions. For a more robust, legal and thoughtful AI landscape in India, that is necessarily the first step.

64 Vidushi Marda, The State's Data Overreach, India Today, December 3 2021, https://www.indiatoday.in/magazine/up-front/story/20211213-the-state-s-data-overreach-1883538-2021-12-03.

65 The Supreme Court e-Committee published its vision document in April 2021, and granted repeated extensions for submission of responses to the Draft. Additionally, the e-Committee invited three civil society members to form a sub Committee to help envision Phase III of the e-Courts project. This is much more than what is usually adopted by AI policy initiatives in the country, although there is still significant scope to make it a more critical, engaging and receptive exercise.

CHAPTER 04

Chinese AI Governance in Transition: Past, Present and Future of Chinese AI Regulation

04

CHAPTER 04

Chinese AI Governance in Transition: Past, Present and Future of Chinese AI Regulation

JULIA CHEN（陈英）

Abstract

This chapter examines how China's approach to domestic AI regulation has developed since the 2017 publication of its New Generation AI Development Plan and considers how it might evolve in the coming few years. It describes how while the period from 2017 to 2020 mostly saw reliance on soft regulation, including self-regulation by AI companies, hard regulation from government actors increased between late 2020 and 2021. Eschewing a purely top-down narrative, it highlights how hard regulation was spurred partly by academics, netizens and media reports. At the same time, it acknowledges the role of policymakers' desires to restrain the "wild growth" of internet platforms and steer resources towards applications of AI perceived as more socially and economically beneficial. The chapter concludes by considering the possible impacts and future directions of Chinese AI regulation.

Any views expressed in this chapter are those of the author and do not necessarily represent the views or positions of any entities she is associated with.

Introduction

China issued its overarching artificial intelligence (AI) strategy, the New Generation AI Development Plan (AIDP), in 2017.[1] Recognizing the major strategic opportunities but also the challenges presented by AI, the plan said that by 2025, China would "have seen the initial establishment of AI laws and regulations, ethical norms and policy systems, and the formation of AI security assessment and control capabilities."[2] Roughly midway through this timeframe, this chapter examines how China's approach to domestic AI regulation has developed since the publication of the plan and considers how it might evolve in the coming years.[3] While regulation can be and has been used to expand or accelerate the adoption of AI technologies, the focus here is largely on actions to keep their development and deployment within acceptable bounds.[4]

This chapter describes how while the period from 2017 to 2020 mostly saw reliance on soft regulation, including self-regulation by AI companies, hard regulation from government actors increased between late 2020 and 2021.[5] Eschewing a purely top-down narrative, it highlights how hard regulation was spurred partly by academics, netizens and media reports. At the same time, it acknowledges the role of policymakers' desires to restrain the "wild growth" of internet platforms and steer resources towards applications of AI perceived as more socially and economically beneficial.[6] The chapter concludes by considering the possible impacts and future directions of Chinese AI regulation.

1 For a detailed English-language analysis, see Huw Roberts et al., "*The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation*," AI & Soc 36 (2021): 59–77, https://doi.org/10.1007/s00146-020-00992-2.

2 "*Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)*," last modified August 1, 2017, https://web.archive.org/web/20220215010101/ https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/.

3 Although there are signs of China's aspirations to contribute to international AI governance, detailed discussion of these aspirations and how well they are being realized is outside the scope of this chapter.

4 As evidence of regulation being used for promoting AI research and adoption, a search for "*artificial intelligence*" (人工智能) in the title field of a widely used legal database, Wolters Kluwer China Law & Reference (威科先行法律信库), on September 14, 2021 found 247 results in the "local laws and regulations" (地方法规) category. These included documents establishing innovation competitions, industrial parks and innovation zones, announcing educational/talent-development measures, and inviting applications for research funding. There are also examples of local governments mandating use of certain AI technologies in specific settings.

5 As recognized in Keith Sisson and Paul Marginson, "'*Soft Regulation*' – Travesty of the Real Thing or New Dimension?", ESRC "*One Europe or Several?*" Programme (2001), there is no widely accepted definition of "soft" or "hard" regulation, and different forms of regulation are arguably best thought of as being arrayed along a continuum running from "soft" to "hard". This chapter defines hard regulation as specific obligations and rights typically enforced by sanctions, and soft regulation as behaviour-guiding frameworks, principles, and standards outside of traditional law.

6 The language of 野蛮生长, which we translate as "wild growth" but has also been translated as "savage growth" or "unfettered growth", was used for instance in a meeting of the Central Committee for Comprehensively Deepening Reform in August 2021. See: "习近平主持召开中央全面深化改革委员会第二十一次会议," cctv.com, published August 30, 2021, https://baijiahao.baidu.com/s?id=1709521220719079200&wfr=spider&for=pc.

Given the huge breadth of applications and sectors that use AI, this analysis does not aim at comprehensiveness. Draft AI legislation issued by the local government in Shenzhen, which could provide clues as to the future direction of national AI regulation, is a major object of focus. In addition, the chapter examines the governance of three AI-based techniques that have been widely applied across industries or emerged in conjunction with new business models: facial recognition technology (FRT), differential pricing, and algorithmic recommendation.[7] While regulation in the last two areas can be seen as part of a crackdown on large consumer-facing platforms that began in late 2020, other aspects of this so-called "techlash" are largely out of scope. There are two reasons for this. First, while AI technologies are core to the operations of many of the targeted platforms, many of the behaviours they have been punished for are not directly related to algorithmic practices. These include forcing exclusivity on sellers and blocking links to each other's services. Second, the techlash has received significant treatment from Western media and analysts, with this chapter attempting to highlight some other areas that have received comparatively little attention.[8]

A difficulty in attempting to write about AI regulation in China is that the landscape is moving quickly, making it possible that by the time of publication, legislation still in draft form at the time of writing in early 2022 will have been confirmed or revised, and other noteworthy developments will have emerged. The authors hope the readers will forgive any inaccuracies and omissions that result.

2017 to 2020: Self-Regulation, Soft Regulation

The three years following the publication of the AIDP have been described by Dr. Xue Lan, Chair of the AI govcernance expert committee under the Ministry of Science and Technology (MOST), as a time of "responsive governance" (回应式治理).[9] The dominant policy aim was driving AI development, while some soft regulation was introduced in response to certain problems.

Hard regulation of AI in this period either targeted narrow applications with clear safety or economic risks or responded to prominent incidents. For instance, regulatory requirements were issued for drone use, automated driving, and AI-

7 This article uses "differential pricing" to refer to what is described in Chinese as 大数据杀熟, literally "big data killing old customers". This is the use of big data analytics to apply different pricing strategies to different users, which can result for instance in higher prices for existing users than for new users.

8 Western media analysis includes for instance: "*SupChina's 'China's 'Big Tech crackdown': A guide',*" SupChina, published August 2, 2021, https://web.archive.org/web/20220222222102/https://supchina.com/2021/08/02/chinas-big-tech-crackdown-a-guide/; and "*China's attack on tech,*" Economist, published August 14, 2021, https://www.economist.com/weeklyedition/2021-08-14.

9 "专访清华大学薛澜：社会治理应尽快进入人工智能敏捷治理阶段", published July 13, 2021, https://www.tsinghua.edu.cn/info/1662/85834.htm.

based asset management services.[10] One of the most high-profile incidents occurred around the end of August 2019 when an app called Zao that allowed users to swap their faces into scenes from popular dramas provoked public backlash about its weak privacy protections.[11] The Ministry of Industry and Information Technology told Zao's parent company to take corrective measures including revising its terms. Zao swiftly removed the offending clause and issued an apology to users.[12] A few months later, the Cyberspace Administration of China (CAC), which is responsible for internet content regulation, issued provisions restricting deepfake transmission.[13]

Generally though, development of AI during this period was guided by softer policy tools. The concept of taking an "accommodating and prudent" (包容审慎) approach to new technologies and business models, promoted by Premier Li Keqiang since 2017, helps explain this decision. The concept was described in a Chinese media article reposted on the national government's website as a wait-and-see approach of trusting the market, encouraging innovation, and strengthening regulation as and after incidents happen (as opposed to ex-ante).[14] Li suggested it would support job creation.[15] The idea was also employed in a white paper on AI governance issued in September 2020 by the China Academy for Information and Communications Technology (CAICT) and China's Artificial Intelligence Industry Association.[16] The white paper stated that an accommodating and prudent approach to AI regulation would help avoid hindering the development of AI. It identified companies as the main AI governance entities in the near term and suggested using industry agreements,

10 Drone use: "*Provisions on the Administration of the Real-name Registration of Civil Unmanned Aircrafts,*" accessed September 30, 2021, https://web.archive.org/web/20220222222310/http://lawinfochina.com/display.aspx?id=24026&lib=law; Automated driving: "三部委关于印发《智能网联汽车道路测试管理规范（试行）》的通知," published April 11, 2018, https://web.archive.org/web/20220222222721/https://www.miit.gov.cn/zwgk/zcwj/wjfb/zbgy/art/2020/art_699ad3bae2bb45759e7a5d39a4073c54.html; AI-based asset management services: "关于规范金融机构资产管理业务的指导意见," accessed September 30, 2021, https://web.archive.org/web/20220222223131/https://baike.baidu.com/item/%E5%85%B3%E4%BA%8E%E8%A7%84%E8%8C%83%E9%87%91%E8%9E%8D%E6%9C%BA%E6%9E%84%E8%B5%84%E4%BA%A7%E7%AE%A1%E7%90%86%E4%B8%9A%E5%8A%A1%E7%9A%84%E6%8C%87%E5%AF%BC%E6%84%8F%E8%A7%81/22458019?fr=aladdin.

11 "*Chinese netizens get privacy-conscious,*" Economist, published September 7, 2019, https://www.economist.com/business/2019/09/07/chinese-netizens-get-privacy-conscious.

12 Runhua Zhao, "*Face-Swap App Zao Apologizes for User 'Confusion' and Promises Data Protection,*" Caixin Global, published September 4, 2019, https://www.caixinglobal.com/2019-09-04/face-swap-app-zao-apologizes-for-user-confusion-and-promises-data-protection-101458603.html.

13 "关于印发《网络音视频信息服务管理规定》的通知," published November 18, 2019, https://web.archive.org/web/20211230143244/http://www.cac.gov.cn/2019-11/29/c_1576561820967678.htm. The rules went into effect on January 1, 2020. In January 2022, CAC issued draft provisions for more comprehensively regulating synthesised content. See: "国家互联网信息办公室关于《互联网信息服务深度合成管理规定（征求意见稿）》公开征求意见的通知", published January 28, 2022, https://web.archive.org/web/20220208062554/http://www.cac.gov.cn/2022-01/28/c_1644970458520968.htm.

14 "所谓包容审慎监管，表面看，就是先看一看、先放一放、先让市场多跑一跑，但这并非权宜之策。从长远看，它是一种基础性理念。这种理念意味着相信市场、鼓励创新，同时更加重视加强事中事后监管。" From: "李克强为何反复强调要秉持"包容审慎"的监管理念？," published September 23, 2017, https://web.archive.org/web/20210816234404/http://www.gov.cn/xinwen/2017-09/23/content_5227149.htm.

15 "李克强总理_明确要求：“各部门都要树立‘包容审慎’监管理念，为就业创造更大空间。" , Ibid.

16 CAICT is a think tank under the Ministry of Industry and Information Technology.

ethical norms, and technical guidance in this early period, before moving to legal constraints once the technology had matured.[17]

AI principles were a prominent component of the soft regulation path taken during this period. High-profile examples were those issued around mid-2019 by a government-sponsored research hub, the Beijing Academy of Artificial Intelligence (BAAI), and MOST's AI governance expert committee.[18] Like many similar documents issued in other countries, however, there was a lack of clarity about how implementation would occur. Although CAICT's white paper described how Chinese companies Baidu, Tencent, and Megvii—along with Western companies—had developed their own AI principles and were attempting to put them into practice, evidence of concrete changes effected by these corporate initiatives is lacking.

Finally, the creation of technical standards by industry actors and academics, coordinated by government bodies and industry associations, has constituted another relatively soft tool for shaping AI development, given their non-mandatory nature. China's approach to standards differs from that of the US, another leader in AI, in its greater centralization as well as its greater use of national standards alongside participation in international standard-setting.[19] Between 2018 and 2021, the number of national AI standards doubled from around 80 to 162.[20]

17 "人工智能治理白皮书," published September 28, 2020, https://web.archive.org/web/20220223002314/http://m.caict.ac.cn/yjcg/202009/t20200928_347546.html.

18 BAAI principles: "《人工智能北京共识》正式发布 北京智源人工智能研究院人工智能伦理与安全研究中心在京揭牌," 机器之心, published May 25, 2019, https://web.archive.org/web/20220225084632/https://www.jiqizhixin.com/articles/2019-05-25-6; AI governance expert committee principles: "Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI'," published June 17, 2019, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai.

19 On the greater centralization of standardization efforts in China, see: Jeffrey Ding, "*Balancing Standards: US and Chinese Strategies for Developing Technical Standards in AI,*" published July 1, 2020, https://www.nbr.org/publication/balancing-standards-u-s-and-chinese-strategies-for-developing-technical-standards-in-ai/. Whereas CESI's AI standardization white papers separately list international and national standards relevant to AI, a plan released by the US National Institute for Standards in Technology (NIST) in August 2019 presented a single table summarizing areas in which AI standards had been developed, with footnotes mostly referring to international standards. See: "*US LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools,*" published August 9, 2019, https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf. NIST's domestic work on AI seems more focused on the production of datasets and other assessment tools and convening stakeholders to address fundamental questions about AI use. See: "*Industries of the Future,*" published January 15, 2020, https://www.nist.gov/speech-testimony/industries-future.

20 "人工智能标准化白皮书（2018版)," published January 24, 2018, https://web.archive.org/web/20220225085123/http://www.cesi.cn/201801/3545.html; "人工智能标准化白皮书（2021版）," published July 19, 2021, https://web.archive.org/web/20220225085326/http://www.cesi.cn/202107/7796.html. For an overview of the structure of the AI standards system, see: "五部门关于印发《国家新一代人工智能标准体系建设指南》的通知", published July 27, 2020, http://www.gov.cn/zhengce/zhengceku/2020-08/09/content_5533454.htm.

Late 2020 to 2021: The Move to Hard Regulation

In late 2020 to 2021, government actors moved towards the use of harder regulatory measures covering certain applications of AI.[21] Dr. Xue has characterized this as a period of "concentrated governance" (集中治理), in which multiple departments took action after problems that had emerged previously did not receive a sufficiently focused or timely response.[22]

Measures taken since the start of this phase include prohibitions on the use of FRT without consent in commercial settings, bans on differential pricing, and regulations on algorithmic recommendation representing the first of their kind globally. Meanwhile Shenzhen's draft Regulations for the Promotion of the AI Industry, issued in July 2021, constitute the country's first cross-sector AI legislation at the local level.[23] They include measures to shape as well as boost AI development and deployment, listing various governance-related responsibilities for different actors. According to the draft, the city government should use innovative regulatory measures, including policy handbooks and regulatory sandboxes, and establish a platform for AI testing and certification and an AI ethics committee. The committee's responsibilities will include tracking companies' implementation of ethical norms and researching, monitoring and providing judgements on major areas such as algorithmic discrimination, deepfakes, and impacts on social structure.[24] "Harder" regulatory measures put forward in the draft include requirements for AI companies to conduct assessments of ethics and safety/security risks and provide staff training on such risks. In addition, high-risk AI applications will be subject to regulatory supervision before market launch. This bears some similarity to the risk-based approach in the AI regulation proposed by the European Union (EU), though criteria for determining risk level are not provided in the Shenzhen draft.[25]

21 This is not to say that there was no continuity from the 2017 to 2020 period. For instance, in the area of standards, CAICT in June 2021 launched an initiative to evaluate FRT systems against desiderata such as security, reliability, transparency, and data protection. See: "护脸计划进入新阶段，可信人脸识别测试正式启动," published June 8, 2021, https://web.archive.org/web/20220223003254/. By January 2022, over 100 providers had been approved under the voluntary scheme. See: "破百！第二批"可信人脸应用守护计划"成员单位名单出炉", published January 21, 2022, https://web.archive.org/web/20220225092427/https://mp.weixin.qq.com/s/z7DU9jrRE5gVBUpNYWZ5uQ.

22 "专访清华大学薛澜：社会治理应尽快进入人工智能敏捷治理阶段", published July 13, 2021, https://www.tsinghua.edu.cn/info/1662/85834.htm. We translate 集中治理 as "concentrated governance," following China Daily: "More responsible, trustworthy technology called for", China Daily, published July 20, 2021, https://global.chinadaily.com.cn/a/202107/20/WS60f62020a310efa1bd662f1a.html.

23 "深圳经济特区人工智能产业促进条例（草案）," published July 14, 2021, https://web.archive.org/web/20220223003515/http://www.szrd.gov.cn/rdyw/fgcayjzj/content/post_713069.html.

24 The meaning of "social structure" in this context is not clear, but it could refer to the potential implications of increased automation on the labour market and wealth distribution.

25 For the European Union's proposed regulation, see "*A European approach to artificial intelligence*," last modified February 23, 2022, https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.

We now turn to the question of why the volume of AI-related hard regulation increased during this period.

Motivations

Responding to Public Discontent

First, much of the regulation addresses concerns raised by academics, netizens, and media outlets, which have grown as applications of AI have proliferated.

Academics have not only supported the development of soft regulation through the standards-setting described above, but have also helped to accelerate hard regulation. The first FRT lawsuit was brought by a legal professor against an Hangzhou wildlife park in October 2019, sparking widespread discussion of the associated privacy risks and highlighting deficiencies in the relevant legislation.[26] The Vice-President of China's Supreme Court mentioned the case as part of his explanation of the court's July 2021 interpretation of FRT.[27] The interpretation stated that businesses must acquire users' consent before collecting facial information, and the provision of products or services cannot be conditioned on consent if facial information is not necessary.[28]

Academics have also helped justify regulatory action through their research, as in the case of differential pricing. In March 2021, a Fudan University team released a ride-hailing study based on booking more than 800 trips. It found that users of more expensive phones were more likely to have their requests received by more expensive cars. In response, the deputy secretary-general of Shanghai's consumer protection committee recommended that relevant regulatory departments take action to avoid consumer interests being harmed.[29]

The attitudes of netizens have also played an important role in prompting regulatory action. The Supreme Court supported its interpretation of FRT by citing a public opinion survey conducted in the first half of 2020 by a task force commissioned by

26 Yuan Ye, "*A Professor, a Zoo, and the Future of Facial Recognition in China*," published April 26, 2021, https://web.archive.org/web/20220217125812/https://www.sixthtone.com/news/1007300/a-professor%2C-a-zoo%2C-and-the-future-of-facial-recognition-in-china.

27 "《最高人民法院关于审理使用人脸识别技术处理个人信息相关民事案件适用法律若干问题的规定》新闻发布会," published July 28, 2021, https://web.archive.org/web/20220223003903/https://www.court.gov.cn/zixun-xiangqing-315831.html.

28 "最高人民法院关于审理使用人脸识别技术处理个人信息相关民事案件适用法律若干问题的规定，" published July 28, 2021, https://web.archive.org/web/20220223004248/https://www.court.gov.cn/fabu-xiangqing-315851.html.

29 "算法来收割，乘客变韭菜：网约车玩大数据"杀熟"2.0版," 半月谈, published April 26, 2021, https://mp.weixin.qq.com/s/JDNfTXiCMKCi0REyKzOUsQ.

four government departments.[30] The survey found that over 64 per cent of more than 20,000 respondents thought that FRT had a tendency to be abused, and over 30 per cent had already suffered loss or privacy infringement through the leakage or abuse of facial data.[31] As well as being channelled through such centrally-driven surveys, complaints have also surfaced on and been amplified by social media. In November 2020, a video went viral of a man wearing a helmet to evade tracking when visiting a real estate exhibition. This led to local market supervision authorities in Zhejiang and Jiangsu fining tens of property sales companies for infringing personality rights using FRT.[32]

Differential pricing, prohibited in various regulations including the Personal Information Protection Law (PIPL) in 2021, had started to attract attention from netizens years earlier.[33] A Zhihu thread launched in May 2018 that asked: "How to evaluate the phenomenon of differential pricing?" had attracted over 1,000 responses and 6.5 million views by September 2021.[34] In some cases, users went a step further and sought legal redress for perceived injustices. In July 2021, following a couple of failed attempts by users to sue superapp Meituan and travel platform Ctrip for differential pricing in the preceding years, a court in the city of Shaoxing became the first to rule in favour of a plaintiff in a case of this kind.[35] The plaintiff had complained

30 "《最高人民法院关于审理使用人脸识别技术处理个人信息相关民事案件适用法律若干问题的规定》新闻发布会," published July 28, 2021, https://web.archive.org/web/20220223003903/https://www.court.gov.cn/zixun-xiangqing-315831.html. For details of the task force, see: "关于APP违法违规收集使用个人信息专项治理工作这有一份详尽的查询指南!", published November 7, 2019, http://www.cac.gov.cn/2019-11/07/c_1574658334765452.htm.

31 The same survey found, however, that the majority response (selected by 65 per cent of respondents) to a question about the value of FRT and how it should be promoted was, "Overall, advantages outweigh the disadvantages. As applications are rolled out it is still necessary to pay attention to risks and guarantee users' rights to be informed and choose." "六成受访者认为人脸识别技术有滥用趋势", published October 19, 2020, https://web.archive.org/web/20220223041338/http://ha.people.com.cn/n2/2020/1019/c351638-34357546.html.

32 "人脸识别400+行政处罚案例分析解读," 来胜教育, https://web.archive.org/web/20220223041455/https://zhuanlan.zhihu.com/p/401307386; Jiayun Feng, "Viral video of man evading facial recognition leads to surveillance bans in Chinese cities," SupChina, published December 3, 2020, https://web.archive.org/web/20220223041759/https://supchina.com/2020/12/03/viral-video-of-man-evading-facial-recognition-leads-to-surveillance-bans-in-chinese-cities/.

33 State Administration for Market Regulation regulations: "市场监管总局关于对《价格违法行为行政处罚规定（修订征求意见稿）》公开征求意见的公告," published July 2, 2021, https://web.archive.org/web/20220223041858/https://www.samr.gov.cn/hd/zjdc/202107/t20210702_332196.html; "市场监管总局关于《禁止网络不正当竞争行为规定（公开征求意见稿）》征求意见的通知," published August 17, 2021, https://web.archive.org/web/20220223041847/http://www.moj.gov.cn/pub/sfbgw/lfyjzj/lflfyjzj/202108/t20210817_434872.html. A provision relating to differential pricing (Article 24) was included in the final draft of PIPL, issued on August 20, 2021, but not the previous two drafts, as shown in: China Briefing Team, "China's Personal Information Protection Law: A Comparison of the First Draft, the Second Draft, and the Final Document", published August 26, 2021, https://www.china-briefing.com/news/chinas-personal-information-protection-law-a-comparison-of-the-first-draft-the-second-draft-and-the-final-document/. For Shenzhen's data and draft AI regulations, which include provisions on differential pricing, see footnotes 49 and 23 respectively.

34 Zhihu is an online forum similar to Quora. "如何评价大数据杀熟这一现象？," accessed September 30, 2021, https://web.archive.org/web/20220223042307/https://www.zhihu.com/question/268104462 .

35 Meituan case: "美团涉"大数据杀熟"被质疑：你以为你在薅羊毛，但你才是被薅的羊，" 搜狐科技官方帐号, published December 18, 2020, https://baijiahao.baidu.com/s?id=1686412041836398824&wfr=spider&for=pc; failed attempt to sue Ctrip: "郑育高与上海携程商务有限公司侵权责任纠纷二审案件二审民事判决书," published February 1, 2021, https://susong.tianyancha.com/6454d6c58071460a808a715d5cebd578.

that Ctrip had sold her a hotel room at an above-market rate based on her high ability to pay. However, the court did not discuss differential pricing in-depth in either the first hearing or the subsequent appeal hearing; damages were ultimately awarded on the basis of fraud and illegal processing of personal data.[36] Analysis by law firm LLinkslaw suggested that this was because of the difficulty of proving differential pricing. In the future, a PIPL provision that has since come into effect, which makes information processors liable for damages if they cannot prove they are not at fault, should make it easier for users to win similar lawsuits.[37]

Investigative reports, often amplified by netizens, have also helped to expose problems with AI applications. A report on the treatment of food delivery workers in September 2020 prompted measures to deter the use of algorithms to exploit workers or put their safety at risk.[38] Among other provisions, guiding opinions jointly released by seven government departments in July 2021 to uphold such workers' rights prohibited the use of "the most stringent algorithm" (最严算法) as an assessment requirement. Instead, they promoted the use of "moderate algorithms" (算法取中).[39] Algorithmic recommendation management provisions (henceforth referred to as the ARMP) issued in draft form in August 2021 and finalized at the end of that year, stated that algorithmic recommendation providers offering work dispatch services to workers must "fulfil the duty to ensure workers' rights and interests".[40]

> "Academics, netizens and journalists have thus all played an important role in prompting local and national regulators to place stronger constraints around AI systems and the data they rely on."

36 Jian Fang and Shenjie Si, "不授权就不给用？绍兴女士状告携程APP，法院判了," published July 8, 2021, https://web.archive.org/web/20220223042800/http://www.shaoxing.com.cn/p/2877040.html.

37 "大数据杀熟第一案代理随笔——简评胡某诉携程案二审判决", published January 2022, http://www.llinkslaw.com/uploadfile/publication/9_1643105020.pdf.

38 Youxuan Lai, "外卖骑手，困在系统里," 人物, published September 8, 2020.

39 "新规|关于落实网络餐饮平台责任切实维护外卖送餐员权益的指导意见," published July 26, 2021, https://web.archive.org/web/20220223043153/https://www.163.com/dy/article/GG5OHLTA0551M1HQ.html. In the absence of an established translation in the English-speaking world, we use "moderate algorithms" for 算法取中.

40 "*Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022,*" published January 10, 2022, https://web.archive.org/web/20220223043219/https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/.

The Drive From Top Leadership To Reduce Wealth Inequality And Ensure Political Stability

At the same time, some of the constraints on AI's application introduced in this period must be understood in the context of hardening attitudes among political leaders towards the role of big tech—and capital more broadly—in the economy and society. Policymakers had become concerned that consumer internet platforms were abusing their market power at the expense of consumers and employees. President Xi Jinping sees reining in these platforms as part of a drive to realize common prosperity or reduce inequality, which he has described as "not just an economic issue but also an important political issue relating to the foundation of the Party's rule."[41] Presiding over a meeting of the Central Committee for Comprehensively Deepening Reform (CCCDR) in August 2021, Xi said that it was necessary to promote fair competition, support small-medium enterprises in particular, and protect consumer rights through high-quality development and promoting common prosperity.[42] Signals from top leadership have thus likely contributed to legislative and enforcement bodies' readiness to take action against differential pricing and the negative impact of algorithms on delivery workers.

Furthermore, concern about the influence of large internet platforms on political stability is evident in the ARMP. The provisions confirm that algorithmic recommendation services that allow the public to express opinions or have social mobilization capabilities must comply with rules introduced in November 2018 requiring internet information services with such properties to conduct a "security assessment" and submit it to the authorities.[43] They also restate an earlier prohibition on synthetic fake news content and certain provisions on the online content ecosystem that took effect in 2020.[44] The latter requires providers to present content conforming to "mainstream value orientations" in prominent areas, take down and report illegal or "undesirable" content, and establish mechanisms for manual intervention and user choice.[45] A new addition, echoing the risk-based approach set out in Shenzhen's draft AI regulations, is the establishment of a graded algorithm security management

41 "习近平：共同富裕是社会主义的本质要求，是中国式现代化的重要特征" 求是网, published August 22, 2021, https://web.archive.org/web/20220223043420/http://www.qstheory.cn/zhuanqu/2021-08/22/c_1127784024.htm. "Common prosperity" was espoused by Xi's predecessors but has become increasingly prominent in 2021; it appeared 65 times in Xi's speeches and meetings as of the end of August, up from 30 in the whole of 2020. "What is China's common-prosperity strategy that calls for an even distribution of wealth?," South China Morning Post, published August 26, 2021, https://www.scmp.com/economy/china-economy/article/3146271/what-chinas-common-prosperity-strategy-calls-even.

42 "习近平主持召开中央全面深化改革委员会第二十一次会议," published August 30, 2021, https://baijiahao.baidu.com/s?id=1709521220719079200&wfr=spider&for=pc. The CCCDR is one of China's most important policymaking bodies.

43 "具有舆论属性或社会动员能力的互联网信息服务安全评估规定," published November 15, 2018, https://web.archive.org/web/20220223043642/http://www.cac.gov.cn/2018-11/15/c_1123716072.htm.

44 For the fake news provision, see Article 11 of: "网络音视频信息服务管理规定", published November 29, 2019, https://www.chinalawtranslate.com/en/provisions-on-the-management-of-online-a-v-information-services/.

45 A translation is available at: "*Provisions on the Governance of the Online Information Content Ecosystem*," published December 21, 2019, https://www.chinalawtranslate.com/en/provisions-on-the-governance-of-the-online-information-content-ecosystem/.

system based on factors such as social mobilization capability and number of users.[46] In consolidating and expanding upon past efforts, the ARMP show that the government recognizes the significant power commanded by recommender services over public opinion and wants to ensure that this power is kept under close supervision.

Promoting Development in Priority Areas While Highlighting Red Lines

Despite the tightening restrictions on certain uses of AI, the government remains convinced of the economic and social opportunities provided by the technology and is keen to promote its development and targeted deployment.

This nuanced position is evident in trailblazing data and AI legislation in Shenzhen. Shenzhen has been able to pursue an ambitious legislative programme in these areas because of its status as a demonstration pilot zone for socialism with Chinese characteristics, building on its history as a pioneer in reform and opening up.[47] In October 2020, in a five-year implementation plan for the pilot zone, the central government expressed its support for Shenzhen to strengthen legislative explorations in emerging fields.[48]

Passed in June 2021, Shenzhen's data regulations include measures to promote the integration and sharing of public data, which could have substantial benefits for the AI industry, while also placing limits on data collection and the deployment of algorithms.[49]

Shenzhen's draft AI regulations, according to an explanatory document issued by the local congress, aim to push forward the implementation of China's national AI strategy while establishing the world-class status of Shenzhen's AI industry.[50] Developed through gathering feedback from over 200 companies of various sizes, universities, and research institutions, they seek to balance the promotion of innovation and the firm protection of redlines.[51] They represent lawmakers' attempts

46 Details of the grading system, including how social mobilization capability will be assessed, have not been made public as of February 2022.

47 China's economic reform and opening to international markets began in 1978. In 1980, Shenzhen was one of four cities to be designated a Special Economic Zone (SEZ). SEZs received a range of economic and political privileges and were encouraged to test new policies and institutions for a market-oriented economy. For more, see Douglas Zhihua Zeng, ed., Building Engines for Growth and Competitiveness in China: Experience with Special Economic Zones and Industrial Clusters (Washington, D.C.: World Bank, 2010).

48 "中共中央办公厅 国务院办公厅印发《深圳建设中国特色社会主义先行示范区综合改革试点实施方案（2020－2025年）》," 新华社, published , November 10, 2020, https://web.archive.org/web/20220220094721/http://www.gov.cn/zhengce/2020-10/11/content_5550408.htm.

49 "《深圳经济特区数据条例》全文公布！" published July 7, 2021, https://web.archive.org/web/20220223043943/http://www.sznews.com/zhuanti/content/2021-07/07/content_24368291.htm.

50 "关于《深圳经济特区人工智能产业促进条例（草案）》的说明," published July 14, 2021, https://web.archive.org/web/20220223003515/http://www.szrd.gov.cn/rdyw/fgcayjzj/content/post_713069.html.

51 The redlines set out in Clause 78 seem to be activities already prohibited by law, such as infringing on personal privacy, providing products and services that harm national or social public security, etc.

to find a middle way between what they describe as Europe's severe and cautious approach to governance and the lax "no need for approval" approach of the US.[52] If these attempts succeed, they may help usher in Dr. Xue's proposed next phase of "agile governance" (敏捷治理), which promotes both innovation and risk mitigation in a coordinated way.[53]

> Developed through gathering feedback from over 200 companies of various sizes, universities, and research institutions, they seek to balance the promotion of innovation and the firm protection of redlines.[51] They represent lawmakers' attempts to find a middle way between what they describe as Europe's severe and cautious approach to governance and the lax "no need for approval" approach of the US.[52] If these attempts succeed, they may help usher in Dr. Xue's proposed next phase of "agile governance" (敏捷治理), which promotes both innovation and risk mitigation in a coordinated way. [53]

Shenzhen's regulations highlight the following areas as priorities for AI adoption: social services (healthcare, education, elderly care, etc), social governance (tax collection, environmental protection, public order, etc), and economic development (manufacturing, financial services, etc). They thus reinforce signals provided in national government strategy documents about the importance of integrating digital technologies such as AI with manufacturing and public services.[54] Combined

52 "关于《深圳经济特区人工智能产业促进条例（草案）》的说明," published July 14, 2021, https://web.archive.org/web/20220223003515/http://www.szrd.gov.cn/rdyw/fgcayjzj/content/post_713069.html.

53 The other features of this mode are: (1) consideration of multiple objects of governance: data, algorithms, application scenarios, companies and platforms; (2) multistakeholder participation; and (3) use of a range of hard and soft tools in a flexible manner. "专访清华大学薛澜：社会治理应尽快进入人工智能敏捷治理阶段", published July 13, 2021, https://www.tsinghua.edu.cn/info/1662/85834.htm.

54 These documents include the AIDP and the Outline of the 14th Five-Year Plan, published in March 2021. The Outline of a Five-Year Plan sets out China's strategic objectives, with hundreds of subsequent sub-plans containing more details about how the objectives will be met. The Outline of the 14th Five-Year Plan calls for the promotion of the deep integration of digital technologies with the real economy. It also lists 10 application settings for digitization: transport, energy, manufacturing, agriculture and irrigation, education, health, culture and tourism, community services, home devices, and government services. See: "中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要," published March 13, 2021, http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.

with stricter controls on consumer-facing recommendation services at the national level, these signals are trying to direct AI talent and funding towards areas that government actors see as more socially and economically beneficial.

Effects and Limitations

Having considered the main motivations behind AI regulation from 2020 to 2021, we now turn to the potential effects.

An optimistic reading of the developments might see them as necessary and timely. Although the legislation of 2020-21 was in some cases tackling issues that had been observable for some time, once regulators decided to act they did not hang around. The high-level nature of Chinese legislation has helped to facilitate speedy action, signalling the direction of travel for companies as they await more detailed regulations and responding to emerging social issues before they become larger problems.[55] By contrast, the EU's proposed legislation on AI is more comprehensive and detailed but is not expected to come into effect until 2024 at the earliest.[56]

A couple of examples of Chinese internet firms responding to government signals rather than waiting for enforcement details are provided by ByteDance and Meituan. The former announced in September 2021 that it was limiting the time spent on its video platform Douyin by under-14s to 40 minutes a day.[57] This followed the issuing of the draft ARMP, which specified that providers "may not use algorithmic recommendation services to lead minors to online addiction", and the imposition on the gaming industry of tightened time limits for minors earlier in the month. Responding to transparency requirements in the PIPL and ARMP, Meituan publicly issued explanations of its arrival time estimation and order dispatch algorithms in September and November 2021 respectively.[58] These examples show how internet companies are adjusting their behaviour to adapt to the tighter regulatory environment.

55 As noted in Jie Tang and Anthony Ward, The Changing Face of Chinese Management (London: Routledge, 2003), "Laws are... often framed in vague and general terms. This has often been necessary to allow their passage in the first place... Vagueness also allows for further experimentation, with details fleshed out in specific regulations later." Similarly, a scholar involved in the drafting of the PIPL said, "Law can't be too rigid in these advanced fields. Therefore, the draft PIPL only sets principles, and its implementation will be at the discretion of relevant supervisory departments, as well as data processors themselves, to explore the benefit of algorithms in terms of personalisation without violating the bottom line of law." See: "*Top Scholar Zhou Hanhua Illuminates 15+ Years of History Behind China's Personal Information Protection Law,*" published June 8, 2021, https://web.archive.org/web/20220223162143/https://digichina.stanford.edu/work/top-scholar-zhou-hanhua-illuminates-15-years-of-history-behind-chinas-personal-information-protection-law/.

56 "*EU proposes new Artificial Intelligence Regulation,*" published April 21, 2021, https://www.nortonrosefulbright.com/en/knowledge/publications/fdfc4c27/eu-to-propose-new-artificial-intelligence-regulation#:~:text=The%20AI%20Regulation%20envisages%20a,apply%20as%20early%20as%202024.

57 Coco Feng, "Chinese version of TikTok limits kids under 14 to 40 minutes per day, adding to fight against internet addiction," South China Morning Post, published September 20, 2021, https://www.scmp.com/tech/policy/article/3149397/chinese-version-tiktok-limits-kids-under-14-40-minutes-day-adding-fight.

58 "让更多声音参与改变，美团外卖"订单分配"算法公开" , published November 5, 2021, https://web.archive.org/web/20220223162638/https://mp.weixin.qq.com/s/qyegF_r_SPGnkEdZqkVjxA.

As for the impact on the wider technology sector, there is some early evidence that regulation may be successfully steering AI funding towards policy priorities, rather than having a dampening effect on the overall innovation landscape. Chinese start-ups managed to attract about a third more venture capital funding in 2021 than 2020, with investments in robotics and semiconductors increasing while funding for e-commerce and social start-ups stayed flat or declined.[59]

On the other hand, some of the legislative provisions may fail to protect consumers as intended. For instance, the requirement in the ARMP to allow users the option to switch off algorithmic recommendations may see limited uptake. One reason is that personalized recommendation is often genuinely valuable for users, helping them more quickly to find content that interests them. Another is that even if users are aware of their right to switch off personalized content and want to exercise it, there may be some friction involved in doing so. For instance, it takes between five and seven steps for users of Douyin, Taobao or Meituan to go from the homepage to turning off personalized content recommendations.

59 Ryan McMorrow, "Investors are shunning China's once-hot consumer tech start-ups," Financial Times, published January 31, 2022, https://www.ft.com/content/25ae00fa-8913-4cdc-ba14-7b011a8c9928.

How UX design on Taobao (left) and Douyin (right) discourages users from opting out of personalized ads and personalized content respectively. Left: "After turning off personalized recommendations, we will not send adverts according to your interests, but you will still see the same number of adverts, and they may be adverts you are not interested in." Right: "After turning off personalized content recommendation, we will show you content based on the popularity of the content and other non-personalised elements. You may see content you are not interested in or don't like, your user experience will be impacted." In both cases, the option on the right is highlighted to encourage users to keep personalization switched on. Source: Screenshots from author's phone, taken September 14, 2021.

The requirement in the same regulations to let users choose or delete user tags used for algorithmic recommendation services may also see limited adoption, if (albeit limited) evidence from outside China is a good predictor. While Facebook users can delete interests and other categories used to show them targeted adverts on their ad preferences page, only 14 per cent of Facebook users surveyed by Pew Research in Autumn 2018 knew about this page.[60]

In addition, a blanket restriction on differential pricing could potentially have unintended consequences. One systematic review of the ethics of algorithmic pricing shows a combination of both material benefits, such as efficiency and choice, and negative impacts, such as a feeling of having been duped, that may be harder to quantify.[61] Another study concludes that "while big data seems poised to revolutionize pricing in practice, it has not altered the underlying principles. In particular, value-based pricing [where prices reflect differences in consumers' willingness to pay for a particular good or service] generally benefits sellers who earn more profit and buyers who would otherwise be priced out of the market, at the expense of less price-sensitive customers who end up paying a higher price."[62] Research into the impact China's ban has on consumers across the income spectrum could help ensure that the net effects are positive and inform the policy approach of other countries.

Finally, it is important to note that the use of AI by the public sector has in general not been prioritized for regulatory oversight in China to the extent that it has in the EU and the US.[63] It is true that Shenzhen's draft AI regulations specify that algorithms relating to public decision-making or public interests should be explained in a comprehensible way, and PIPL regulates government as well as private sector collection of data. However, the latter maintains broad authority for state organs to handle personal information, and facial recognition by law enforcement will still be

60 Emily A. Vogels, "The longer and more often people use Facebook, the more ad preferences the site lists about them," published December 3, 2019, https://www.pewresearch.org/fact-tank/2019/12/03/facebook-ad-preferences-linked-to-frequency-of-use-age-of-account/. It is possible that the proportion is now higher, but we could not find more up-to-date data as of September 2021.

61 Peter Seele et al., "*Mapping the Ethicality of Algorithmic Pricing: A Review of Dynamic and Personalized Pricing*," Journal of Business Ethics 170 (2021): 697–719.

62 "*Big Data and Differential Pricing,*" 19, published February, 2015, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf.

63 The EU's proposed regulatory framework for AI classes the use of AI for essential public services and law enforcement as high-risk, thereby subjecting such applications to stringent requirements. See "*Regulatory framework proposal on Artificial Intelligence*," accessed September 30, 2021, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Various state governments in the US have drafted legislation on automated decision-making by government entities. See: "*Legislation Related to Artificial Intelligence*," modified September 15, 2021, https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx.

permitted.[64] By contrast, the proposed EU regulation on AI only allows for real-time FRT to be used in a few narrowly defined contexts.[65]

Moreover, as the Chinese government seeks to steer the deployment of AI towards socially beneficial applications, the potential remains for well-meaning interventions in these areas to exacerbate pre-existing issues or create new harms. Taking the use of AI in education as an example, some Chinese academics have begun to discuss the potential ethical challenges, but these seem under-explored in government discourse.[66]

Summing up this section, China's AI regulations have so far prioritized speed and responsiveness over detail and comprehensiveness. While this approach may have benefits given the pace with which the technology is advancing, clarification and iteration will be needed in the years to come.

Where Next?

Acknowledging the ongoing nature of this endeavour, nine government departments issued guiding opinions, in September 2021, setting out their intention to spend the next three years or so gradually establishing an integrated governance system for internet information service algorithms.[67] This will include working out the details of governance mechanisms announced in the ARMP. In the process, companies will be trying to negotiate rules that limit disruption to their operations as far as possible. A report on explainable AI issued by Tencent shortly after the confirmation of the ARMP argued that "as experience abroad shows, legislation should avoid adopting overly restrictive regulatory requirements and making one-size-fits-all requirements for the application of AI algorithms in terms of transparency and interpretability."[68]

How might Shenzhen's AI regulations influence the broader trajectory of AI governance? According to a researcher at a Shanghai-based science and technology

64 For discussion of the applicability of PIPL to Chinese state organs, see Alexa Lee et al., "*Seven Major Changes in China's Finalized Personal Information Protection Law*," published September 15, 2021, https://digichina.stanford.edu/work/seven-major-changes-in-chinas-finalized-personal-information-protection-law/ .

65 "*Proposal for a Regulation laying down harmonised rules on artificial intelligence*," accessed February 25, 2022, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

66 On ethical challenges, see "教育人工智能伦理风险如何消解" , published April 28, 2021, https://web.archive.org/web/20220223163230/https://theory.gmw.cn/2021-04/08/content_34747804.htm. A Ministry of Education notice about progress in deploying AI to support teachers do not discuss ethical risks: "教育部关于实施第二批人工智能助推教师队伍建设行动试点工作的通知", published September 8, 2021, https://web.archive.org/web/20220223163301/http://www.moe.gov.cn/srcsite/A10/s7034/202109/t20210915_563278.html.

67 "关于印发《关于加强互联网信息服务算法综合治理的指导意见》的通知", published September 17, 2021, https://web.archive.org/web/20220223163349/http://www.gov.cn/zhengce/zhengceku/2021-09/30/content_5640398.htm.

68 Tencent, "Explainable AI Development Report 2022: Concepts and Practices for Opening the Black Box of Algorithms", accessed February 24, 2022, https://docs.qq.com/pdf/DSmVSRHhBeFd0b3Zu. For an English excerpt translated by Daniel Zhang, see: https://docs.google.com/document/d/1LtTkZRF5SNqDTkOqg1mPCaWzb6YbQwHfg4zQQuOhrUo/edit#.

think tank: "While the regulations are not detailed, and a lot of discussions—including with other cities—would be needed before adapting them to the national setting, they provide a valuable starting point for discussion and improvement." Only a handful of cities and areas would have the legislative authority required to imitate the regulations.[69] Many localities are instead taking the approach of monitoring the impacts of AI through surveys and other experimental methods. In cities and provinces including Beijing, Zhejiang, Guangdong and Hubei, academia, industry and government are cooperating to study the impact of AI in different settings and collate case studies to learn from.[70]

This reflects a broader trend of including a wider range of stakeholders in AI governance—one of the desired features of Dr. Xue's proposed "agile governance" mode. For instance, the ARMP encourages industry associations to supervise and guide service providers in establishing and implementing standards and complying with the law.[71] Shanghai's trade unions have proposed a greater role for mechanisms allowing negotiation between workers and platforms in formulating algorithms.[72] Following the reference to AI ethics committees in Shenzhen's draft AI regulations, MOST guiding opinions on strengthening science and technology (S&T) ethics stated that institutions engaged in sensitive research in fields such as AI should establish ethics review committees.[73] The close relationship between the state and other actors mentioned here may bring both advantages (such as lending authority to those

69 One area that could be increasingly important in AI regulation is Shanghai Pudong New Area, which was given special legislative powers in June 2021. Later that year, Shanghai's data regulations stated that Pudong New Area would establish an algorithmic assessment standards system and drive Intellectual Property protection for algorithms. "上海市数据条例", published November 29, 2021, https://web.archive.org/web/20220209143137/https://www.shanghai.gov.cn/nw12344/20211129/a1a38c3dfe8b4f8f8fcba5e79fbe9251.html.

70 For more on these "social experiments", see Jun Su, "Steadily Taking Off: China's AI Social Experiment Is in Full Swing," in Shanghai Institute for Science of Science, "*AI GOVERNANCE IN 2020*," published June 2021, https://www.aigovernancereview.com.

71 Article 5, "*Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022*," published January 10, 2022, https://web.archive.org/web/20220223043219/https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/. The China Computer Federation's Committee on Ethics and Professional Conduct, established in 2020, is an example of an effort in this direction. See: "CCF又有大动作——揭秘职业伦理委员会," published September 18, 2020, https://www.ccf.org.cn/Chapters/CEPC/news/2020-09-18/729522.shtml.

72 "制定平台算法，“小哥”们的声音不应缺失", published January 18, 2022, https://web.archive.org/web/20220223163824/http://www.news.cn/fortune/2022-01/18/c_1128272488.htm.

73 The opinions also commit to promptly elevating important S&T ethical norms to laws and regulations. The final text, approved by the CCCDR in December 2021, has not been released as of mid-February 2021. For the draft text, see: "科技部公开征求对《关于加强科技伦理治理的指导意见（征求意见稿）》意见," published July 28, 2021, https://web.archive.org/web/20220223164006/https://mp.weixin.qq.com/s/YCo206biPpgO4zTo3gBiSA. According to Zheng Liang, Vice-Dean of the Institute for AI International Governance of Tsinghua University, the regulations are targeted at S&T research and development rather than commercial activity. See: Yaning Li, " 科技部：人工智能企业或需设立科技伦理审查委员会," published July 31, 2021, https://web.archive.org/web/20220223164127/https://www.sohu.com/a/480669649_161795. While the establishment of ethical norms was mentioned in the 2017 AI Development Plan, the guiding opinions can also be seen as a response to the gene-editing of babies by Jiankui He in 2018, which prompted domestic and international criticism.

actors' activities) and limitations (for example on their scope to explore public sector misuse of AI). [74]

Finally, we highlight an area of AI ethics that has not yet garnered significant attention in China but may attract more discussion in future. Many of China's top AI labs are pushing the boundaries of large pre-trained language models, which have applications for a broad range of downstream tasks including chatbots and text generation. Yet among several publications by Chinese institutions introducing such models in 2021, there was limited discussion of their potential to reinforce bias or be misused, despite these risks receiving growing attention in international research.[75] One possible reason is that whereas prominent incidents in the West have led more researchers there to focus on algorithmic bias, similar incidents have been less common in China, or at least not publicized as widely.[76] Another possible reason is a concern among Chinese companies that discussing downside risks without presenting solutions may attract unwanted attention from the media or regulators. Whatever the explanation, researchers and regulatory bodies—including any new ethics committees—may need to pay more attention to these risks as language models with powerful generalizing abilities are developed.

74 For instance, all trade unions formed under an employer in China must be registered with the All-China Federation for Trade Unions, a government body. China's industry alliance model is also considerably state-driven. For more on industry alliances, see: Xiangzi Chen and Liping Guo, "*Analysis on the Lessons Drawn from International Experience to Promote China's Industry Alliance innovation*", Modern Economic Information, no. 1 (January 2019): 385.

75 Chinese models released in 2021 include CPM, developed by Tsinghua University and BAAI researchers, Pangu by Huawei and RecurrentAI researchers, and M6 by Alibaba and Tsinghua researchers. Exceptions that did discuss ethical challenges include: Ming Ding et al., "*Cogview: Mastering text-to-image generation via transformers*," Advances in Neural Information Processing Systems 34 (2021), https://arxiv.org/pdf/2105.13290.pdf; and Shaohua Wu et al., "Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning," arXiv (2021), https://arxiv.org/pdf/2110.04725.pdf. An industry report by BAAI on large intelligent models also acknowledged ethical and environmental challenges posed by such models: Tiejun Huang et al., "超大规模智能模型产业发展报告," 62, published September 2021, available at: https://web.archive.org/web/20220225110643/https://hub.baai.ac.cn/view/10263. For an example of international discussions of risks from "foundation models" including large pre-trained language models, see parts 5.1 and 5.2 of Rishi Bommasani, Percy Liang et al., "*On the Opportunities and Risks of Foundation Models,*" arXiv (2021), https://arxiv.org/pdf/2108.07258.pdf.

76 Examples in the West include the labelling of black people as gorillas by Google Photos ("Google apologises for Photos app's racist blunder", BBC, published July 1, 2015, https://www.bbc.com/news/technology-33347866) and an automated hiring algorithm used by Amazon that disadvantaged female candidates ("*Why Amazon's Automated Hiring Tool Discriminated Against Women*", published October 12, 2018, https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against.)

In summary, the period from 2017 to 2021 has seen a shift towards stricter regulation of how AI can be used in China. This shift has occurred in response to growing public concerns about privacy, security and fairness, combined with a desire from political leaders to rein in the "wild growth" of large tech platforms while promoting the deployment of AI in more strategic areas.[77] The pace of regulatory activity does not look likely to let up in the short term, as implementation details will need to be worked out and remaining gaps identified and addressed. Throughout, government authorities, drawing on the expertise and experience of other sectors of society, will be trying to suppress prominent abuses of the technology, while realizing the potentially huge economic and social benefits of AI. If they succeed, elements of their approach could well be adopted elsewhere.

77 See footnote 6 for more context on the term "wild growth" (野蛮生长).

CHAPTER 05

The Myth of Data-Driven Authoritarianism in Asia

05

CHAPTER 05

The Myth of Data-Driven Authoritarianism in Asia

CINDY LIN AND YUCHEN CHEN

Abstract

Artificial intelligence (AI) in Asia is not simply a duel between superpowers and the production of top-down, state-controlled technology. In this chapter, we show how AI systems in Asia are produced through careful negotiations of familial and kin relations between state and society. Moreover, AI and data-driven technologies are not solely mandated by the state to control and regulate citizens but are also a place to work out the future of politics in the region of Asia. In this chapter, we show how citizens and junior government engineers retool AI and data-driven technologies to both contest and leverage long-standing "Asian values" and familial relations held between state and society to achieve democratic ideals and address social problems. Through long-term ethnographic fieldwork and archival research in Indonesia and China, we argue that AI and data-driven technologies are not simply tools to enact authoritarian governance in Asia as often-depicted in Western media but also techniques to intervene in oppressive social, political, and economic conditions and ideologies. We propose methodological recommendations that move beyond typecasting Asian societies as democratic and/or authoritarian in order to seek out a "situated ethics" that can regulate the negative implications of a data-driven world.

Introduction

Across Western media outlets and scholarly literature, it is common to find stories of non-Western governments committed to acts of surveillance and control with the use of data-driven and artificial intelligence (AI) technologies.[1] Historical descriptions in popular Western discourses of the Social Credit System (SCS) paint China's government as ruthless leaders implementing Orwellian surveillance tools against citizens.[2] In Indonesia, the rollout of AI in governance has been used to surveil labour and manage logistics,[3] prompting public commentary that data-driven technology has infringed on individual rights.[4] In both scenarios, AI and data-driven technology are pictured either in Western media or by foreign policy experts as key nodes for authoritarian control.[5] To that end, data- and AI-driven technology and authoritarianism in Asia have become sutured to one another in ways that foreclose how citizens can relate to the state otherwise—or even how the state itself functions to control those within it. Accounting for state priorities, institutional contexts, and cultural norms around governance in Asia, our chapter aims to complicate simplistic links made between AI and authoritarian control.

Our chapter challenges the tightly bounded relationship between AI and data-driven technology and authoritarianism through archival and ethnographic research in Indonesia and China. Far from being a tool of surveillance and control, we show how citizens and junior government engineers retool AI and data-driven technologies to

1 Feldstein, Steven. The global expansion of AI surveillance. Vol. 17. Washington, DC: Carnegie Endowment for International Peace, 2019. https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847; Andersen, Ross. "*The Panopticon is Already Here.*" The Atlantic, September 2020. https://www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/; Goode, Kayla and Heeu Millie Kim. "*Indonesia's AI Promise in Perspective.*" (Center for Security and Emerging Technology, August 2021). https://doi.org/10.51593/2021CA001

2 See, for example, Canales, Katie. "China's 'Social Credit' System Ranks Citizens and Punishes Them with Throttled Internet Speeds and Flight Bans If the Communist Party Deems Them Untrustworthy." Business Insider. Business Insider, December 24, 2021. https://www.businessinsider.com/china-social-credit-system-punishments-and-rewards-explained-2018-4.

3 Grill, Gabriel. "*Future Protest Made Risky: Examining Social Media Based Civil Unrest Prediction Research and Products.*" Computer Supported Cooperative Work (CSCW), 2021. https://doi.org/10.1007/s10606-021-09409-0.

4 Fikri, Aulia. "*Between Risk, Trust, and Habit: A Survey on Unicorn Companies Application Users in Indonesia on their Level of Awareness and Concern towards Big Data and Data Privacy.*" PhD diss., (Ritsumeikan Asia Pacific University, 2021); Goode and Kim, 2021.

5 See, for example, Powers-Riggs, Aidan. "*Covid-19 Is Proving a Boon for Digital Authoritarianism.*" Covid-19 is Proving a Boon for Digital Authoritarianism | Center for Strategic and International Studies, July 30, 2021. https://www.csis.org/blogs/new-perspectives-asia/covid-19-proving-boon-digital-authoritarianism. And Heath, Ryan. "*China's Tech Authoritarianism Too Big to Contain.*" POLITICO,, November 20, 2020. https://www.politico.com/news/2020/11/20/chinas-tech-authoritarianism-438646.

both contest and leverage long-standing "Asian values"[6] and familial relations held between state and society to achieve democratic ideals and address social problems.

> In other words, AI in Asia is not simply a technology for totalitarian control over society; rather they have been mobilized and campaigned by its users for contesting, negotiating, and even resisting authoritarian values and/or for ideals of social justice.

In China, unlike popular portrayals of SCS as a tool for the Chinese state to police citizens—and conversely write a surveillance history of a top-down state-controlled system—our chapter demonstrates how the system's beginnings were initiated by Chinese citizens eager to achieve modernization and social justice, especially in tackling problems related to industry fraudulent practices and bad debts. Moreover, the citizen-driven effort was made possible because of how the state perceives its own responsibility to citizens and subsequently organized itself accordingly: the parent as the state and children as citizens.

In Indonesia, the use of agile software development methods,[7] typically associated in the West as a means to control and surveil one's labour and productivity,[8] was central for how junior engineers in a state engineering agency reinvented long-standing social hierarchies within governance. Junior engineers use agile software development frameworks to contest values such as hierarchy and communitarianism that have long been necessary for authoritarian leaders in Indonesia to thrive.
In other words, AI in Asia is not simply a technology for totalitarian control over society; rather they have been mobilized and campaigned by its users for contesting, negotiating, and even resisting authoritarian values and/or for ideals of social justice.

6 Authoritarian values in Asia typically centre around the "*Asian values*" debate in the 1990s where political leaders from China, Indonesia, Singapore, and Malaysia promote social harmony, cooperation, and communitarianism. These countries used such values to argue that liberal democracy from the West, and in particular the United States, is not appropriate for Asia. Consider how in 2004, Indonesia's ex-vice president Jusuf Kalla claimed that Asian values of cooperation rather than competition is necessary for the country to survive, even as it grapples with the debt of the Asian financial crisis, and an increasing income inequality.

7 Set in contrast to an older "waterfall" model of software development that is fairly top-down, long, and requires heavy documentation, agile software development methods promote self-management among team members. Members are tasked to review work processes while completing them on short time-scales, making iterative and adaptive software development core to agile software development methods.

8 Moore, Phoebe V. "*Tracking affective labour for agility in the quantified workplace.*" Body & Society 24, no. 3 (2018): 39-67.; Bjørn, Pernille, Anne-Marie Søderberg, and S. Krishna. "*Translocality in global software development: The dark side of global agile.*" Human–Computer Interaction 34, no. 2 (2019): 174-203.

In the following two case studies, we show how AI and data-driven technologies have been enrolled and deployed to contest authoritarian structures, ranging from deep-seated social hierarchies in governance to the initiation of a system to promote trustworthiness and achieve social harmony. While we acknowledge that these two case studies might not generalize across regions or even within the nation itself, our chapter aims to provide an in-depth and lived account of how such AI technologies and its affiliated software development methods are designed and implemented in practice. This includes paying attention to the technology's lifeline; how it was conceived, legitimated, and circulated. More crucially, our chapter does not start with Western, liberal values such as privacy and transparency as the primary destination for AI ethics, but begins with the cultural norms and political histories of two of the world's largest countries to articulate the importance of a situated AI ethics.

Indonesia / CINDY LIN

Indonesia's transition away from its authoritarian regime is often lauded by foreign policy experts as a miracle. Before 1998, the world's largest Muslim democracy suffered from the dictatorship of President Suharto, a leader that endorsed military power to suppress dissidents and manage state-owned enterprises. The 2019 reelection of Joko Widodo, the first president without military affiliations or an elite background, represented what many called "a break from the past".[9] This break is most manifest in Widodo's invitation to Nadiem Makarim, CEO of decacorn ride-share app Gojek, to join his presidential cabinet, leading many experts to conclude that Indonesia is headed towards a brighter, tech-oriented, democratic future.[10]

Widodo's project to make Indonesian bureaucracy "productive" and market-competitive is motivated by long-held political commitments to reform an "older" paternal patronage system, one that dates back to Indonesia's New Order (1966-1998). The New Order was led by Suharto who cultivated a hegemonic form of Javanese masculinity called "bapak". Bapak established the norms for what it meant to be a modern, male leader in Indonesia.[11] Bapakism (directly translated as "fatherism") confined women to domestic arenas and served as an authority to younger men, ensuring that social stratifications registered the rule of the male authority as well as the harmony of the familial structure—or as Suharto called

9 "*Indonesia's Successful Democratic Transition Adds New Momentum to US-Indonesian Relations,*" U.S. Embassy & Consulates in Indonesia, November 3, 2014, https://id.usembassy.gov/indonesias-successful-democratic-transition-adds-new-momentum-to-us-indonesian-relations/.

10 Prabowo was appointed as the new Minister of Defense, and Nadiem as the Minister for Education and Culture, with Widodo reasoning that Nadiem can bring about democratic reforms and innovation to an outdated education system.

11 Boellstorff, Tom. "*The Emergence of Political Homophobia in Indonesia: Masculinity and National Belonging.*" Ethnos 69, no. 4 (December 1, 2004): 465–86, https://doi.org/10.1080/0014184042000302308.

himself, "the father of the nation".[12] Materially, this paternalism was rooted in the ways in which Suharto offered concessions to his own banks and domestic monopolies—who in turn offered his rule political support.

In this section, I will document how the implementation of agile computing methods and technologies in government research agencies used to develop AI systems have become infused with a new ideological importance within post-colonial governance. These methods promise to transform the cultural norms and subjectivities of Indonesian bureaucrats—officials who have historically been portrayed by international media and development agencies as paternalistic and corrupted. Based on long-term ethnographic fieldwork with Indonesia's first engineering agency, Agency for Application and Assessment of Technology (the Badan Pengkajian Penerapan Teknologi or BPPT), and their partnership with IBM, I show how junior male engineers in BPPT not only adopt IBM's agile computing methods, but also associate them with Indonesia's reform ideals of transparency and accountability introduced during the country's democratic transition in 1998.[13] In this way, junior male engineers appropriated corporate tools typically associated with neo-liberal techniques of self-improvement as a means to intervene in broader social hierarchies and paternalistic structures of power (Bapakism).

The IBM Garage Methodology (hereon referred to as IBM Garage), a consultancy platform that introduces clients to agile methods, design thinking, and DevOps (an integration of software development and IT Operations) was first contracted by BPPT in October 2019. BPPT hired IBM to develop the agency's capacities in AI as the new "AI Innovation Centre". The director of the agency contracted IBM's services to lead a digital transformation programme that would teach BPPT how to think like an agile start-up. To change the mindset of civil servants, IBM introduced 15-minute daily stand-up meetings to BPPT, a typical method found in agile software development cycles. Over a period of three months, IBM and BPPT members would meet each other in virtual stand-up meetings to update each other about what they did yesterday, what they intended to do on that day, and if they faced any obstacles.

12 Nilan, Pam. "*Contemporary Masculinities and Young Men in Indonesia*," Indonesia and the Malay World 37, no. 109 (November 1, 2009): 327–44, https://doi.org/10.1080/13639810903269318.; Intan Paramaditha, "*Contesting Indonesian Nationalism and Masculinity in Cinema: Ingenta Connect*," Asian Cinema, Volume 18, Number 2, September 2007, pp. 41-61(21); Julia I. Suryakusuma. State Ibuism: The social construction of womanhood in the Indonesian new order. Institute of Social Studies, 1988.

13 These ideals were first introduced to Indonesia in 2001 by pro-democracy Reformasi (Reform) activists at the very time these young male engineers entered university, where calls and protests demanding for reform in governance were common. Aspirations for embedding tech entrepreneurship within bureaucratic life—and endorsing agile computational techniques as effective bureaucratic practice—are outcomes of the popularization of discourse of transparency and accountability in political and cultural spheres following Indonesia's transition from authoritarianism to democracy in 1998. It follows then, that young Indonesian engineers, mostly men, in BPPT are adopting IBM's agile development practices, to remake the bureaucracy in the vein of principles of transparency and accountability that gained traction during Indonesia's reform period.

This exercise also required that participants quantify in advance how long they would take to finish a given task, such as programming a front-end interface in JavaScript. The IBM Garage team, which is composed of 13 designers, developers, and cloud architects from IBM Asia Pacific, led the sessions by commending those who have done their tasks before their deadlines. Many of the junior engineers employed by BPPT turned to these techniques to differentiate their contributions from the older employees. This included the demonstration of speed and delivery of immediacy measured through finishing tasks within a set period of time.

To render oneself productive and transparent meant to show precisely how many hours one had worked, which was different from the bimonthly or trimonthly meetings BPPT used to organize. Infrequent meetings had created a product backlog so huge that junior engineers ended up completing what senior engineers had not efficiently delegated. By adopting methods of agile software development, junior engineers hoped to transform BPPT finally into the kind of productive and useful state agency the Widodo presidency had envisioned, exactly because these processes enabled the quantification of their work productivity.

Evoking Family in Bureaucracy

Over the next few weeks, junior engineers from BPPT continued to attend the stand-up meetings, instead of the full team. Repeatedly, IBM Garage leaders expressed disappointment about the lack of attendance. While IBM Garage members initially viewed the poor attendance as evidence of the work ethic of senior engineers, they eventually reorganized the work to include junior engineers as leaders of back-end and front-end software development. While I initially thought that senior engineers would feel left out of the development process, Adika,[14] a senior engineer, explained why he didn't think so.

He reasoned that BPPT is a family, and social hierarchies between senior and junior engineers are natural. To him, junior engineers should work to ensure the continued welfare of BPPT. Being in a family justified the lack of involvement from senior engineers in the development of the system. I probed further, asking why junior engineers had to do the work of senior engineers. He looked at me, bewildered, before teasing me that my understanding of working in BPPT was shaped by my own upbringing in a modern city-state like Singapore, where there were no "rituals" or "traditions".

14 All names have been pseudonymously used.

> “BPPT is like a family. If Dayuf [a senior engineer] wants to ask anything or ask for help, he should be able to ask junior members. Everyone doesn’t have to do something. There will be some people who don’t want to contribute, and there will be some members who want to contribute but are not brave enough to make a website... Look at Rajah (junior engineer), he was assigned to do something whereas I volunteered my time. I chose to do it, whereas some others were asked to do something that they cannot say no to.”

By treating your colleagues as family, Adika evoked the figure of the senior engineer as a person of authority or Bapak. For instance, Dayuf, in this family structure, could demand the time of junior engineers, if needed. Adika also set himself in contrast with junior engineer Rajah because he chose where he could allocate his time to. Rajah could not because he was junior in ranking.

In these moments, Adika, like other senior engineers, embodied the state ideology of Bapakism. Dayuf and Adika could demand the time and labour of junior engineers because they embodied the figure of a senior male and father. Such hegemonic forms of masculinity, pervasive during the time of Suharto’s leadership, still found their way into the contemporary practice of agile software development techniques. The Bapak always ruled over the family, even as younger members of the family increasingly questioned the ways in which the family organized itself.

Junior engineers explained why the older generation continued to enforce strict hierarchy. One of them named Budi prescribed it as “bureaucracy is the older generation; flexibility is absent in the older generation.” Budi defined this older generation as senior staff members who entered BPPT under the leadership of German-trained engineer B.J. Habibie, who exercised paternalistic rule over junior engineers.[15] “When I am in the office, I will have to attend a discussion or be pulled by seniors into a chit-chat or presentation or to do things for others—all of these are unproductive for me.” Budi’s refusal to chat or do favours for seniors revealed his search for a new labour ethic. He explained, “I get tired of being asked to do many things in order to get something back. I prefer to be independent and choose how I spend my own time.” It was Budi’s desire for a new kind of work style and mindset—distinct from senior engineers and their associations with a familial and paternalistic style of leadership—that led him to believe that agile methods were necessary for producing a new generation of engineers.

15 Amir, Sulfikar. The technological state in Indonesia: The co-constitution of high technology and authoritarian politics. Routledge, 2012.

Most central to why senior engineers enacted familial hierarchical structures was the influence organicist ideas had over them. First promoted by Dutch-trained legal scholar Dr. Supomo during the New Order, organicist ideas were believed to be inherent to how villagers organized themselves in society. Supomo grounded his claims in adat ("customary") law, an artefact catalogued by Dutch anthropologists who obsessed over Indonesians as harmonious, static, and familial-like.[16] Supomo even argued that traders, politicians, and civil servants place the interests of their own families and friends above themselves.[17] Military lawyers from the 1960s used Supomo's ideas of hierarchy, harmony, and order to advocate that there was no sense of separation between the rulers and the ruled. More specifically, military lawyers and the New Order regime advocated for a state that is organized like a family, influenced as they were by Dutch legal thinking that individuals cannot be separated from society.

Most recently, the AI masterplan that BPPT developed in 2020 codified "pancasila", a state ideology that was advanced extensively by President Suharto, to be the guiding principle and regulation for ethical AI. Pancasila or the Five Principles is an Indonesian ideological system based partially on "purportedly authentic Indonesian values of harmony, co-operation, and the family, and which eschew conflict".[18] During the New Order regime, Suharto had elevated the pancasila ideology to the level of the "sole ideology" of the nation, making it a fundamental tool for propelling authoritarianism and cementing the role of the Bapak as the leader of the family state.

The acceptance of organicist and pancilisa ideas as part of Indonesia's legal system worked against the democratically arranged bureaucracy that junior engineers vied for.[19] Furthermore, the New Order could declare the family state as part of being "Asian" and so fend off criticism of its human rights record and the liberal language

16 Adat laws were first made by Supomo's professor, Dr. van Vollenhoven, who is a pioneer of legal anthropology as well as a theorist of international, constitutional, and administrative law. He supervised at least 67 PhD dissertations on adat and other aspects of colonial law. Adat, Put simply, is traditional laws and customs in Indonesia that had been catalogued and categorized by anthropologists, among others, in the late 19th and early 20th century. His visits to Indonesia cataloguing such "traditional" behaviours have become fundamental to colonial policy. He viewed the people of Indonesia as having one single race based on his ethnographic research in Indonesia and studying other ethnographic accounts. He portrayed Indonesia as a community consisted of a whole, in which all its relationship are organic and all individuals in them are static, balanced and harmonious. He was also the one who advocated for indigenous communities to have a degree of sovereignty over their land, even if the study of adat itself was mostly conceived by Dutch anthropologists. Yet, this recognition of local land rights was not necessarily good for those who were subjected to them. It reinforced the power of "reactionary local elites and rendered local communities in many cases even more vulnerable to outside intrusion" (See more in D. Lev, 'Colonial Law and the Genesis of the Indonesian State', Indonesia, 40, (1985): 64, 66).

17 R. Supomo, Hubungan Individu dan Masjarakat dalam Hukum Adat (first published 1941), Jakarta: Pradnya Paramita, 16.

18 Ramage, Douglas E. Politics in Indonesia: Democracy, Islam and the ideology of tolerance. (London, UK: Routledge, 2002).; Bowen, John R. "*On the political construction of tradition: Gotong Royong in Indonesia.*" The Journal of Asian Studies 45, no. 3 (1986): 545-561

19 Bourchier, David. Illiberal democracy in Indonesia: The ideology of the family state. Routledge, 2014.

it was rooted in.[20] I have shown how Adika, a senior engineer who was raised during the New Order regime, viewed his designation of work to junior engineers as part of being a family. Moreover, he saw the use of agile software development methods as destroying such familial relations—relations that were ultimately rooted in the hegemonic masculinity of Bapakism.[21] At the same time, junior engineers aspired to agile methods as a means to intervene in such paternalistic and arguably long-standing organicist ideas.

It follows then that the discursive imaginary of AI is not simply as the West has portrayed, a means for continued authoritarian ruling or as the current Indonesian administration claims, a means for agile governance. Instead, by performing agile work, junior engineers aimed to recreate new kinds of hierarchy that challenged familial relations in bureaucracy. Additionally, the kind of governance structures they endorsed were believed to transform bureaucracies into equitable, transparent, and entrepreneurial organizations.

It is these qualities, as junior engineers have asserted, that would enable Indonesia to become modern and democratic. In this way, it is not easy to ascribe AI and data-driven technologies as central for how non-Western governance enact authoritarianism; such ethics are being re-negotiated at the very level of a stand-up meeting.

“Instead, by performing agile work, junior engineers aimed to recreate new kinds of hierarchy that challenged familial relations in bureaucracy. Additionally, the kind of governance structures they endorsed were believed to transform bureaucracies into equitable, transparent, and entrepreneurial organizations.”

20 Bourchier. Illiberal Democracy in Indonesia, 213

21 Here, what I define as hegemonic masculinity draws from scholarship that recognizes that there is a hegemony of men rather than masculinity itself. The hegemony of men is to “address the double complexity that men are both a social category formed by the gender system and dominant collective and individual agents of social practices” in p. 18, Jeff Hearn and Centre of Gender Excellence, eds., GEXcel Work in Progress Report Autumn 2008 Vol. 5, Vol. 5, (Linköping; Örebro: Institute of Thematic Gender Studies, Department of Gender Studies, Linköping University ; Centre for Feminist Social Studies, Örebro University, 2009).

China / YUCHEN CHEN

In Western media narratives, China's Social Credit System (SCS) has often been cited as an exemplar of dystopian "Orwellian" techno-authoritarianism.[22] However, in detailed documentation of its implementation in Rongcheng, a pilot city of the SCS, a 62-year-old resident comments: "Life in our village has always been good... After introducing the (Social Credit) system, it's gotten even better."[23] This section stays with this tension, one that claims the SCS to be "oppressive" and "unethical" by the West and the other being supportive of such a system on the ground.[24] In this section, I uncover a brief history of the present, of how the SCS is initially envisioned and designed for a market regulation system that is promoted by individual, academic, industry, and governmental efforts and further rolled out as state agenda to address social problems and inequalities against the backdrop of structural changes brought by the market economy in the 1990s.

I read these historical moments with a larger context of paternalism in the governance in China and tease out the harmony ideal crucial for maintaining a familial relationship between the Chinese state as the parent-figure and its citizens. Paternalism is not only a form of governance performed by the leadership, but is also "expected by people from below" as a Confucian ethos.[25] This section shows that this relationship enables citizens to negotiate certain needs instead of passively receiving the state agendas and this relationship enables the state to sustain legitimacy beyond relying on economic development or coercive enforcement.[26] The rhetorics and practices toward harmony and paternalism are what we want to highlight throughout this chapter. I also show the entanglement between China and the West, the state, engineers, and citizens, and academia and industry in co-producing the SCS systems. The entangled relationships further blur the categories and binaries of authoritarianism/democracy and complicate dominant narratives of socio-technical systems in authoritarian states, one that falls into the "authoritarian determinism and reductionism".[27]

The idea of SCS was seeded in 1999. Huang Wenyun, a businesswoman from Shenzhen, wrote a letter to Zhu Rongji, China's prime minister of the time. In the

22 Mosher, Steven W. "*China's New 'Social Credit System' Is a Dystopian Nightmare.*" New York Post, May 20, 2019. https://nypost.com/2019/05/18/chinas-new-social-credit-system-turns-orwells-1984-into-reality/.

23 Mistreanu, Simina. "*China Is Implementing a Massive Plan to Rank Its Citizens, and Many of Them Want in.*" Foreign Policy, April 3, 2018. https://foreignpolicy.com/2018/04/03/life-inside-chinas-social-credit-laboratory/.

24 The high approval of the SCS in Chinese public opinion can also be found in Kostka, Genia. "*China's Social Credit Systems and Public Opinion: Explaining High Levels of Approval.*" New Media & Society 21, no. 7 (2019): 1565–93. https://doi.org/10.1177/1461444819826402.

25 Wang, Wilfred Yang. "*Digital Media, State's Legitimacy and Chinese Paternalism.*" Rowman and Littlefield International, March 13, 2020. https://www.rowmaninternational.com/blog/digital-media-states-legitimacy-and-chinese-paternalism.

26 Wang, 2020.

27 Guan, Tianru. "*The 'Authoritarian Determinism' and Reductionisms in China-Focused Political Communication Studies.*" Media, Culture & Society 41, no. 5 (2019): 738–50. https://doi.org/10.1177/0163443719831184.

letter, she mentioned how the international toy trading business that she built from scratch was badly hurt by pirates.[28] In 1999, China had been opening up the market for 21 years. Yet the market and the state were haunted by countless scandals of Intellectual Property (IP) infringement, product safety issues, and fraud. "Made-in-China" not only connotated cheap labour but also fakeness and bad quality. Huang urged the prime minister to consider the establishment of a nationwide credit system, one "just like the United States". She believed that a credit system was incremental to solve the market problems.[29] Receiving great attention from the central government, Huang followed up with another letter to the Associate President of People's Bank of China and wrote to him with great enthusiasm to contribute to the economic development and national prosperity together and leave some legacy for the later generations.[30] Then, to build that legacy, Huang provided 300,000 RMB (36,200 USD) in 1999 as a research grant for the Institute of World Economics and Politics at the Chinese Academy of Social Science.[31] The funded project aimed to research on how to establish a credit system in China.

Lin Junyue was one of the researchers involved in this project and later became one of the foundational theorists and designers of the SCS.[32] In 1999, Lin had just returned to China after graduating with a Master's in Information Science from Pennsylvania State University. With his expertise in information and economics, he was hired by the Chinese Academy of Social Sciences and went to the United States and Europe for investigation. Bringing back credit-related legal documents and hundreds of hours of interview recordings from the West, Lin went on and built his theory. In 1999 and 2003, Lin published the National Credit System with co-authors and then the Principle of Social Credit System by himself. These foundational texts envisage the core mechanisms, rationalities, and moral foundations of Social Credit. The books were largely built on empirical and theoretical studies of pre-existing Western credit systems, policies, and economic and market theories. However, in his reasoning about the moral gains from introducing credit systems, Lin justifies by citing Zhang Anyuan, a famous economist in China, that, "After all, mores and trustworthiness are nothing but empty faith. Only through the whip of instant karma can we impose those [mores and trustworthiness] on people." Lin then continues to note that the "punishment of

28 Du, Yan. "我为'诚信'五上书[*Five Letters for Trustworthiness*]." Guangming Daily, August 27, 2012.

29 ibid.

30 Ma, Peigui. "黄闻云中国信用体系建设第一人 [*Huang Wenyun: the First Person for China's Credit System*]." Shenzhen Special Zone Daily, September 9, 2012.

31 Ma, 2012. Also in Lin, Yuejun. "林钧跃: 为什么说社会信用体系建设起始于1999年?" [Lin Yuejun: *Why Do We Say the Construction of the Social Credit Started in 1999*?] Credit China, Octover 19, 2019. https://www.creditchina.gov.cn/xinyongyanjiu/xinyongyanjiujiaodianwenzhang/201910/t20191021_172722.html.

32 Besides Lin, there are other scholars working on developing theories of the SCS who were funded by Huang. For example, Wu Jingmei published《现代信用学》[Studies on Contemporary Credits], another foundational text book of credit management. Wu is part of the Credit Management Research Center of Renmin University of China and a drafting group member of the foundational policy text "*China's social credit system construction planning (2014-2020)*."

those who break trust is our whip, and practitioners in credit management are the whip makers."[33]

According to Zhang's logic and Lin's interpretation, to punish those who break the trust is to instil the ideal of harmony into society—a task that can be performed through the technical operations and calculations of the SCS. As Wang points out, "Chinese paternalistic view places great emphasis on the state's role in cultivating its people intellectually and morally, in order to guide them into learning and fulfilling their social obligations and morality".[34] To that end, SCS enforces that all individuals and entities are held accountable and trustworthy by the state. To make the harmony less so an "empty faith", the implementers of the SCS have to ensure the actualization of these qualities into practices. Characterizing the policymakers and practitioners that build such a system as "whip makers", Lin positions the state and its officials as a parent-like disciplinary figure that supervises and disciplines the citizens through moral policing.

> "The SCS should not be merely understood as a controlling system, but also as an "ethical machine" for addressing social problems."

Although the normative and moral standards have been subtly present since the beginning of theorizing China's credit systems, the state still prioritized economic agenda for its ruling legitimacy and as its foremost paternalistic responsibility. In the 2001 symposium organized by the Subcommittee of Economy of National Committee of the Chinese People's Political Consultative (CCPPC),[35] the chair addressed the "practical relevance" of credit systems in China, all of which concerns its economic significance. He argued that credit systems could expand domestic demands by creating more (loan and financial) opportunities for middle- and small-sized enterprises and entrepreneurs and help China "integrate with the world", becoming a member of the World Trade Organization and attracting more investments from

33 Lin, Junyue. 社会信用体系原理 [Principles of Social Credit System]. Beijing: China Fangzheng Press, 2003. [Original text: "说到底，道德和诚信本身只是空洞的信仰，只有通过鞭子和现世报应才能强加到人们的头上……失信惩罚机制是一条'鞭子'，而信用管理行业的从业者就是制造鞭子的人"。]

34 Wang, 2020.

35 The 2001 CCPPC is considered one of the most important symposiums that consolidated the research and pushed the administrative establishment of infrastructure in China's credit system. Small and private businesses and companies, credit and information industry representatives such as Shanghai Credit Information Services (one of the first companies assessing credit scores for individuals starting from 1999), key governmental agencies such as People's Bank of China, academics from Peking Universities and more, gathered together to discuss credit's economic and social conditions in China. The conference also largely drew from legislations in the United States to talk about issues related to privacy, fairness, banking, and more.

the West.[36] In the first official mention of Social Credit, President Jiang Zemin stated China must establish a Social Credit System "compatible with a modern market economy". [37] The urge for a credit system was further legitimized by the Center of Credit Research in the National Reform and Development Commission, that in 2003, around 600 billion RMB (5% of the GDP) was lost due to dishonest market conduct such as fakery and deadbeats.[38]

Such a call for the prosperity of the national economy happened against the backdrop of China's larger shift from the planned economy to the market economy from the 1990s to the early 2000s. During this period, stimulating consumption, (re) allocating and circulating financial resources, and introducing investment were prioritized as a national agenda of modernization. Credit systems during this period were taken up as a must-have tool for economic development that solved the trust issue in the market economy and achieved international legibility and standardization. While there was a national agenda for China to participate in regulated economic development, the state feared that this economic transition would further bring about a lack of trustworthiness and disrupt social harmony.[39] Meanwhile domestically, the early 2000s witnessed increasing grassroots petitions and online activism happening across China.[40] In 2004, President Hu Jintao proposed "harmonious society" in the five-year plan, which has become one of the primary governing goals for the Chinese government since. By "harmonious society", the president meant to have a "social justice guarantee system" that focused on "equal rights, equal opportunities, fair rules and fair distribution".[41] Harmonious society can be read through Communication Scholar Guobin Yang's lens, as a co-evolving "soft" governing practice in reaction to crisis and social unrest.[42]

36 《诚信为本-建立社会信用制度于信用体系研讨会文集》[Trustworthiness as the Foundation - Proceedings of the Establishment Social Credit System Symposium]. Aviation industry press, 2001.

37 Creemers, Rogier. "*China's Social Credit System: An Evolving Practice of Control.*" SSRN Electronic Journal, 2018. https://doi.org/10.2139/ssrn.3175792.

38 Addressed by Xinnian Chen, the chair of the center at NRDC in Shan, Xiuqiao. "陈新年：中国如果没有信用制度GDP总额减少20% [Chen Xinnian: *Without a Credit System, China's GDP Would Decrease by 20%*]." Sohu.com, April 15, 2014. https://business.sohu.com/20060415/n242820483.shtml.

39 Ouyang, Kang. "*Risks in Adopting Modernization as the Way to Build a Harmonious Society in Modern China.*" Essay. In Governance for Harmony in Asia and Beyond, 1st ed., 124–37. Taylor and Francis, 2009.

40 As observed by Sociologist Sun Liping's works about social upheavals and stability in China, see for example, Sun, Liping. "机制与逻辑：关于中国社会稳定的研究" Carnegie Endowment of International Peace. Accessed October 19, 2021. https://carnegieendowment.org/files/sunpaper.pdf.

41 Ouyang, 124.; and as stated in the 16th Central Committee of the Chinese Communist Party, the first proposal and discussion of the "*harmonious society*" campaign, that "构建社会主义和谐社会是一个不断化解社会矛盾的持续过程" [Constructing a socialist harmonious society is a continuous process to solve social conflicts], showing the governing responses as a reaction to social instability. In this key document, the conflicts were framed as the economic development through market economy and social issues that affect harmony, including urbanization, education, trustworthy issues in individuals and government officials, corruption, and more. For the full document, see: http://www.gov.cn/gongbao/content/2006/content_453176.htm.

42 Yang, Guobin. The Power of the Internet in China: Citizen Activism Online. New York: Columbia University Press, 2011.

The "harmonious society" campaign shows the state's response to (re)balance market efficiency and social justice.[43] While deepening economic development, the Chinese Communist Party (CCP) promised to prioritize social harmony, which was considered to be disturbed by unfair distribution of resources and lack of trustworthiness. Given that the Chinese state regards the market economy as fundamentally an imported Western concept, with underlying incommensurable values with "traditional Chinese moral values",[44] action was required to reconcile Western credit and market economy design with Chinese norms of harmony and trustworthiness. I argue that in the state's subsequent development of SCS, various stakeholders, including the state, local governments implementing pilot projects, and key academics, have aligned and deepened the political project of building the harmonious society into SCS, slowly transforming, introducing and codifying the "traditional values" of harmony into the system, so much so that a market device born in the West—and its associations with liberalism—is no longer regarded as a threat and can be adapted in China. As such, the government's ideology of harmony, which emphasizes on "an honest and caring society, and a stable, vigorous and orderly society",[45] is translated into the SCS as a culture of trustworthiness within modern Chinese governance. In the official debut blueprint of the SCS in 2014, the State Council further expanded the concept of "Social Credit" from the financial realm to the socio-cultural and moral realm to include civil judgment, intellectual property, environmental protection, food and drug safety, and more.[46] Moreover, the Council linked SCS and its encoded emphasis on the traditional virtue and moral foundation (of trustworthiness) for national prosperity (under market economy):

> [I]ts inherent requirements are establishing the idea of a sincerity culture and promoting honesty and traditional virtues, it uses encouragement for trustworthiness and constraints against untrustworthiness as incentive mechanisms, and its objective is raising the sincerity consciousness and credit levels of the entire society.[47]

I have shown in this section through the emergence of the Social Credit that ideals of harmony and order are embedded into principles, methods, and tools of governing the social. Moreover, the top-down paternalism performed by the state and mass mobilization of political and social campaigns often work together as shown above and also in other national biosecurity projects in modern China.[48] The

43 In the policy documents, talks, and public discourses of "harmonious society" campaign, the wording of social justice had an emphasis on "公平" [fairness], which emphasized on economic and social (re)distribution and called for rule of law and culture of trustworthiness.

44 Ouyang, 2009.

45 Tao, Julia, Anthony B.L. Cheung, Martin Painter, and Chenyang Li. "*Why Governance for Harmony?*" Essay. In Governance for Harmony in Asia and Beyond, 1st ed., 1–11. London: Routledge, 2010.

46 Jiang, Min. "*A Brief Prehistory of China's Social Credit System*." Communication and the Public 5, no. 3-4 (2020): 93–98. https://doi.org/10.1177/2057047320959856.

47 State Council, State Council Notice concerning Issuance of the Planning Outline for the Establishment of a Social Credit System (2014-2020), 2014.

48 Greenhalgh, Susan. Just One Child: Science and Policy in Deng's China. Berkeley: Univ. of California Press, 2008.

spokesperson(s) for the social credit has been shifting, with multiple voices of key individuals, academics, multiple government agencies, and private sectors involved across stages, from the initiation from the ground to the intervention by the state as a response to the needs and social problems. Therefore, instead of an antagonistic position between the state and citizens, it is one of discipline, but also of negotiation, reactivity, and even reciprocity. The parenthood of the state shifts in its manifestation and positioning in reaction to its ideals and assumptions of its relationship to citizens, one that is about paternalistic give and take. In this local paternalism view, the SCS should not be merely understood as a controlling system, but also as an "ethical machine" for addressing social problems.[49]

Conclusion: Intervening in the Myth of Data-Driven Authoritarianism

Western media and scholarship have focused on the relationship between so-called surveillance capitalism, the rise of authoritarianism and the downfall of Western democracies with renewed concern, especially with the advent of big data and AI technologies in the last two decades. Scholars writing in this vein have looked to the non-West to point out the perils of AI and data-driven technology, arguing that such technologies have furthered both old and new modes of violence and control on class and racial minorities.[50] Our chapter recognizes that this is a possibility, but also complicates the naturalized link between technologies and authoritarian modes of control. Consider how junior engineers in BPPT deploy agile software development methods to intervene in long-standing paternalistic and hierarchical relationships that were crucial for authoritarian leadership during the New Order. Even if agile software development methods typical in AI and data science production have been critiqued as a blueprint for labour exploitation and surveillance, the use and association with democratic ideals such as transparency and accountability show that AI and data-driven technology is ideologically unstable. The various shifts in SCS show how the state, experts, and citizens together react to China's structural changes and its problems brought by the market economy reform under the name of creating harmonious and stable social conditions.[51] Fundamentally, at the heart of SCS is a

49 Ong, Aihwa. "*Introduction: An Analytics of Biotechnology and Ethics at Multiple Scales.*" In Asian Biotech: Ethics and Communities of Fate. Duke University Press, 2010.

50 For example, read Jack, Margaret C., Sopheak Chann, Steven J. Jackson, and Nicola Dell. "*Networked Authoritarianism at the Edge: The Digital and Political Transitions of Cambodian Village Officials.*" Proceedings of the ACM on Human-Computer Interaction 5, no. CSCW1 (2021): 1-25.; Lim, Merlyna. "*The Internet, social networks, and reform in Indonesia.*" Contesting media power: Alternative media in a networked world (2003): 273-288.

51 The shift is not without a political agenda of social control, as other scholars have argued. See, for example, Liang, Fan, Vishnupriya Das, Nadiya Kostyuk, and Muzammil M. Hussain. "*Constructing a Data-driven Society: China's Social Credit System as a State Surveillance Infrastructure.*" Policy & Internet 10, no. 4 (2018): 415–53. https://doi.org/10.1002/poi3.183; and Creemers, Rogier. "*China's Social Credit System: An Evolving Practice of Control.*" SSRN Electronic Journal, 2018. https://doi.org/10.2139/ssrn.3175792.

form of care as control,[52] where paternalistic responsibility initially used by citizens to mobilize social justice slowly transformed into a nationwide credit system to monitor and morally police citizens' everyday lives and behaviours. Simply conceiving AI and data-driven technologies and software development frameworks as key sources of authoritarian ruling and total surveillance in Asia fail to account for the aspirations, desires, and counter-narratives in the race of global supremacy in AI.

In sum, our two case studies have responded to two main calls of this series on AI policy and development in Asia: first, they have expanded the dominant narratives of AI production in the non-West as authoritarian and surveillance-oriented and second; they have captured the responses of both state and non-state actors towards the deployment of such technologies on the ground.

Our chapter described in detail how simply conceiving AI and data-driven technologies and software development frameworks as key sources of authoritarian ruling and total surveillance in Asia fail to account for the aspirations, desires, and counter-narratives in the race of global supremacy in AI. Instead of simply viewing China as a site for illiberal values and Indonesia as a country that has failed to succeed in democratic transformation, we find it necessary to show different trajectories of modernity in these countries and their relationship to technology that far exceeded the democratic versus authoritarian divide. In this way, AI and data-driven technologies have no intrinsic political values and ideologies. Rather, they are technopolitical devices that offer powerful ways to shape countries politically, economically, socially, and culturally.

More importantly, we aim to draw out an ethics of AI that pertains to the ideologies and relations of kinship, harmony, and paternalism crucial for statecraft and subject-making in Asia. Both Indonesia and China have advocated for state ideologies that instituted a seamless relationship between the ruler and the ruled, with paternalism as key for governance. In Indonesia, for instance, organicist ideals were introduced by Dutch-trained legal scholars to institute familial relations between state and society, normalizing the authority of senior male officials over younger men and women. This in turn informed how junior male officials regarded agile software development techniques, key tools for developing AI products today, as revolutionary for transforming governance structures. In this way, the acceptance of AI and data-driven tools in Indonesia is propelled by a desire to intervene in the parochialism of paternalistic relations in governance, rather than simply a nation's race towards AI innovation and leadership. In China, we also see how paternalistic ideologies and the

52 Feminist STS scholars and surveillance scholars have been looking into the dialectical relationship between care and control/surveillance, and calling for attention to the politics of care to see how care functions in practice and affectively for the purpose of control and maintenance of power differentials. See, Martin, Aryn, Natasha Myers, and Ana Viseu. "*The Politics of Care in Technoscience.*" Social Studies of Science 45, no. 5 (2015): 625–41. https://doi.org/10.1177/0306312715602073; for politics of care in the context of East Asia, see Kim, Youngrim, Yuchen Chen, and Fan Liang. *"Engineering Care in Pandemic Technogovernance: The Politics of Care in China and South Korea's COVID-19 Tracking Apps.*" New Media & Society, 2021, 146144482110207. https://doi.org/10.1177/14614448211020752.

ideal of harmony are embedded into the languages, mechanisms, and intentions of the SCS, while not losing the attention to the co-construction of the system by citizens, experts, and the state. We argue that paternalism in governance not only acts through top-down coercive control over its subjects, but is also attentive to the needs from the bottom-up, wherein citizens leverage existing state-society relations and the state can continue to maintain its legitimacy. The historicization of SCS shows a more nuanced and situated story of the import and development of socio-technical systems in China, which is beyond the agenda of authoritarian control.

> “Simply conceiving AI and data-driven technologies and software development frameworks as key sources of authoritarian ruling and total surveillance in Asia fail to account for the aspirations, desires, and counter-narratives in the race of global supremacy in AI.”

In conclusion, the grounds on which governments and public interest bodies use to construct policies and regulation of AI need interrogation, contextualization, and historicization. We have two recommendations to make here with regard to AI governance, both as a way to sidestep normative commitments of AI ethics and to propose a methodological approach to uncovering AI ethics in Asia.

First, AI ethics have been constructed under Western-bound regulatory frameworks, leaving the meanings of “accountability, ethics, and fairness” intact and not open to questioning. In a recent critique of ethical AI/ML, Greene, Hoffman, and Stark argue that high-profile value statements developed by technical experts, multinational tech corporations, intergovernmental organizations and national governments set the “moral background” for what AI ethics is in the first place.[53] This staging of what actions, processes, and values are considered ethical is also an outcome of “technification”: where AI ethics becomes a terrain that depends on the “expert authority” of computer engineers in order to be considered legitimate.[54] This prevents the public from participating in the “democratic oversight” and “intervention” of AI technologies to further equitable and just outcomes for all. We will go a step further

53 Greene, Daniel, Lauren Hoffmann, Anna,and Stark, Luke. “*Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning.*” Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019. https://doi.org/10.24251/hicss.2019.258.

54 Hansen, Lene, and Helen Nissenbaum. “*Digital disaster, cyber security, and the Copenhagen School.*” International studies quarterly 53, no. 4 (2009): 1164 cited in Stark, Luke, Daniel Greene, and Anna Lauren Hoffmann. “*Critical Perspectives on Governance Mechanisms for AI/ML Systems.*” In The Cultural Life of Machine Learning, pp. 257-280. Palgrave Macmillan, Cham, 2021.

> “First, AI ethics have been constructed under Western-bound regulatory frameworks, leaving the meanings of “accountability, ethics, and fairness” intact and not open to questioning.”

to suggest that democratizing intervention into AI technologies is not a given—how can the moral background of AI ethics be formulated and articulated in countries that are “newly” democratic or seeped in authoritarian structures such as Indonesia and China?

Second, historicization of the emergence, mutation, and stabilization of socio-technical systems helps us land on different and/or more expansive understandings of ethics. Historical processes, as historian of computing Mar Hicks notes, reveal the contingency of technological development and question the inevitability of how particular technologies become encoded with certain political values.[55] A historicist sensibility counters presentism in the study of technical systems by attending to “complex and emergent situations over time”.[56] Historicization, in the case of China’s SCS specifically, makes visible how different social actors imbue multiple meanings in the process of constructing, designing, and deploying a system over time. Therefore, historicization helps unpack the fluidity, multivalences and situatedness of ethics. This methodological sensibility to study ethics as it develops in situ challenges the centrality of Western ethical frameworks and its implicit role in AI ethics.

Here, we take the lead of anthropologist Aihwa Ong’s “situated ethics” to consider how context and situation are crucial for thinking about how such a moral background can be formulated.[57] To do situated ethics is to take on its culturally specific meaning especially as they are being negotiated, contested, and lived in relation to others[58] such as at the scale of kin groups, ethnic groups, and the nation. When informed by this standpoint, we are able to adopt a methodological sensibility that locates provisional and situated “ethical configurations” not only

55 Hicks, Marie. “*Hacking the Cis-TEM.*” IEEE Annals of the History of Computing 41, no. 1 (2019): 20–33. https://doi.org/10.1109/mahc.2019.2897667.

56 Robert Soden, David Ribes, Seyram Avle, Will Sutherland. 2021. Time for Historicism in CSCW: An Invitation. In Proceedings of the ACM on Human-Computer Interaction, Vol. 5, CSCW2, Article 459, October 2021.

57 Ong, Aihwa. “*Introduction: An Analytics of Biotechnology and Ethics at Multiple Scales.*” In Asian Biotech: Ethics and Communities of Fate. Duke University Press, 2010.

58 This departs from classical reasoning of ethics of a “self-forming individual” or “to the quest to find a rational form of acting with respect to the good”. Situated ethics allows one to consider how “contemporary problems of living stand in flexible, provisional, and tense interrelationship.” In Ong, Aihwa, and Stephen J. Collier. Global Assemblages: Technology, Politics, and Ethics as Anthropological Problems. Oxford: Blackwell, 2010, 29.

through the individual but also through the ways in which larger collectivities[59] such as how Chinese citizens react to the state's formalization of the SCS or when junior engineers practice democratic values with agile software development methods. In such situated encounters between collectivities, ethical reasoning AI cannot be decided a priori, and any kind of AI policy and regulation will also need to be provisional, flexible, and capable of dealing with cultural complexity. Our chapter opens up the possibility of accounting for the situated, minute practices of resistance, state cooperation with citizens, and bottom-up appropriation. By starting with the everyday lives of citizens and lower-level government officials, we hope to provide a more culturally and historically situated way to approach the project of developing AI governance in Asia.

59 Ong refers to these collectivities as ''communities of fate'': "the network of collectivities that become connected as a result of diverse ethical decisions and feelings associated with technological innovations." In Ong, Aihwa. "Introduction: An Analytics of Biotechnology and Ethics at Multiple Scales." In Asian Biotech: Ethics and Communities of Fate. Duke University Press, 2010, 20.

CHAPTER 06

Kampong Ethics

06

CHAPTER 06

Kampong Ethics

MARK FINDLAY AND WILLOW WONG

Abstract

Our analysis begins with the proposal that the kampong (village) spirit of solidarity, woven into the Asian community identity, can redirect the collaborative applications of artificial intelligence (AI) ethics formulations to maximize AI governance for social good. Accepting Cotterrell's vision of community as social relationships of trust, we argue that locating AI within communities (and thereby the decision-making process underpinning AI design and regulation) creates life-spaces that foster harmonious AI-human coexistence. By rejecting mainstream notions that monolithic/universalist ethical frameworks (often imported from the Western knowledge sphere) can address the ethical priorities of diverse communities across the world, this analysis further argues the "Asian" mode of decision-making and governance is not one specific operation, but whichever that can effectively stimulate and maintain shared trust in the recipient communities. In this analysis, kampong ethics bridges perceived/actual differences in people across external divides, in order to stimulate closer ties of cooperation for diverse communities to collaboratively formulate how to co-exist in harmony with each other and with the tools we create for each other.

Digital Futures Lab | Konrad-Adenauer-Stiftung

Introduction

Starting from the understanding that the "kampong"[1] (village) spirit embodies the strength of healthy reciprocal bonds in a socially located community, this chapter investigates what it means for artificial intelligence (AI) technologies and human agents to exist in a mutually supportive life-space.[2,3,4] In so doing, we will employ Cotterrell's theorizing of community as built on social relations of "mutual interpersonal trust".[5,6] Unfolding in the analysis to follow, the proposed kampong

1 The word "kampong" (otherwise spelled as "kampung", or "kampoeng" in older Indonesian spelling as influenced by Dutch orthography of the Indonesian language) is a Malay world to describe rural village settlements most commonly seen in Southeast Asian countries such as Malaysia, Indonesia, and Singapore (see "*What Is A Kampong?*" 2021. World Atlas. Accessed May 25. https://www.worldatlas.com/articles/what-is-a-kampong-and-where-are-they-found.html.) For the purposes of this research, "kampong" is used as a cultural shorthand for village spirit, wherein benefits are shared, and business is an etiquette of inclusion. In kampong thinking, the world around us is a living, learning institution and new ideas complement that wider world. The notion of gearing ethical decision-making towards, and inspired by the spirit of togetherness and solidarity rooted in a social community applies to geographies outside of the straits. There is no time here and indeed it is not essential for what follows to unpack the kampong myths and challenge the patriarchy, misogyny and mafias which were said to proliferate in some of these lifestyles (see Pan Jie. 2018. "*Can't we just let the Kampung Spirit Die in Peace?*", https://www.ricemedia.co/wp-content/cache/wp-rocket/www.ricemedia.co/rice-media-can-we-let-kampung-spirit-die-in-peace/index.html_gzip).

2 In his account of the social meaning and values attributed to "kampungism", Fausto Barlocco's (2010) suggests the kampung spirit emerges from "a group of people not only residing in an administrative unit, but also having a sense of common belonging based on the sharing of some common practices" (p.405). This description of community spirit resonates with Cotterrell's view of community adopted in this analysis; where our analysis may diverge from Barlocco's position, however, is in his strong emphasis on "the autonomy of the village and its grounding in local practices and traditions" as resistant towards integrations with the regional, national, and global economic and political system (p.405), "in the same way in which its proponents deny theirs on urban salaried labour" (p.422). See Barlocco, Fausto. 2010. "*The Village as a 'Community of Practice*' Constitution of Village Belonging through Leisure Sociality". Bijdragen Tot de Taal-, Land- En Volkenkunde/ Journal of the Humanities and Social Sciences of Southeast Asia 166 (4). Brill: 404–425. Available at: doi:10.1163/22134379-90003609.

3 As no two subjects' experiences are structured the same way, Eric Thompson (2002) explores the various contestations of the feeling around kampong as a "dynamic structure, historically emergent with many uneven edges" (p.55). See Thompson, Eric C. 2002. "*Migrant Subjectivities and Narratives of the Kampung in Malaysia*". SOJOURN: Journal of Social Issues in Southeast Asia 17 (1). ISEAS- Yusof Ishak Institute: 52–75. Available at: doi:10.1355/SJ17-1C.

4 Melani Budianta (2019) contextualizes the spirit of kampung solidarity within the "larger fast-paced urban transformation process", where the "community" formed by socially-bonded people serves as "a strategic site to capture the contradictions, contestations, and local-global cultural dynamics in the Global South" (p.242). This paper reveals that the sense of a "place-based" community is a much-debated concept, as it is "not a given entity", but rather something best described as "a process of becoming that requires effort and work" (p.247). See Budianta, Melani. 2019. "Smart Kampung: Doing Cultural Studies in the Global South". Communication and Critical/Cultural Studies 16 (3): 241–256. doi:10.1080/14791420.2019.1650194.

5 Cotterrell, Roger. 2018. Law, Emotion and Affective Community. SSRN Scholarly Paper ID 3212860. Available at: https://papers.ssrn.com/abstract=3212860, p.2.

6 This analysis also draws heavily from Findlay, Mark, and Willow Wong. 2021. Trust and Regulation: An Analysis of Emotion. SMU Centre for AI & Data Governance Research Paper No. 05/2021, Available at: doi:10.2139/ssrn.3857447.

ethics has three distinct applications that go beyond simple ascription to "things Asian" or to the exclusive primacy of community in AI governance:

- **The** kampong style of communal organization speaks to the preferred nature of ethics formulation in this analysis, where crucial decisions about the creation and deployment of AI are made in service of the community's welfare and actively involve the inputs of community members.
- **By** embedding the kampong spirit in this proposed approach towards AI ethics, the function and position of AI technologies become clarified. In this view, AI lives in the recipient community and so the decision-making process which underpins AI design and regulation is similarly located within a community setting.
- **In** exerting their influence on the process of social bonding in the community, AI ethics and AI technologies in turn are shaped by the requirement to maintain and strengthen the communal relationships of shared trust. Consequentially, the decision-making process must take into consideration the active and dynamic nature of human relationships within a social environment (whether in virtual or physical contexts).

The central concern of this analysis is not the cultural history of the kampong, but the sense of common belonging woven into a community identity, Asian in origin, which continues to underpin the lived experiences of many people within this geographical region called "Asia". As such, adopting this interpretation of the kampong spirit offers a distinct character to the overarching theme of communality advanced in this analysis for the purpose of AI ethics.

Drawing heavily from the theme of communality, the analysis of harmonious AI-human social relations from an Asian context (read alongside its potential implications for the wider global discussions on AI ethics and AI governance formulations) seeks to actively resist further tokenization of what is considered "Asian" or "Western" culture. Efforts in locating or enforcing certain theoretical or empirical-based boundaries to identify certain "cultures" often neglect the reality of diverse communities as built on vibrant collections of constantly evolving and deeply subjective human experiences

which constitute both the “local” and “global” at once.[7] Insert into such a mono reading of cultural diversity (often a shorthand for hegemonic division) a uni-plastic approach to AI-assisted technology and the use of big data, and any comparative or holistic analysis is retarded from reality. A less tokenized understanding of culture requires leaning away from seeing any given community as either monolithic, frozen in time, or neatly packaged in quantifiable and measurable terms from the perspective—supposedly objective and impartial—of the external observer.

The “Asian” contexts referred to in this analysis point to the influence of communal obligations and the concept of the communal “self”, which operate on a significantly more prominent level in Asian societies; whereas, the individualist understanding of liberty, freedom, autonomy and rights is more apparent in Western societies. Returning to earlier points that the “global” inevitably bleeds into the “local” (as will the virtual into the physical), the spirit of communality can be found in Western thinking, and the concepts of freedom and rights are also gaining momentum in Asian societies. Accepting that we all live contemporaneously in a global and local cultural community, universal human rights retain a binding individualist interpretation. For the theme of communality to act as a broad philosophical identifier of cultural predisposition, rather than geospatial location, this analysis leans towards the more inclusive and power-levelling idea of “universal brotherhood/sisterhood”[8] while recognizing the often patriarchal/hierarchical communal frames in Asian social organization.

Consistent with the goal to move beyond classifying and labelling systems of thinking along the overly simplistic dichotomy of “Eastern” and “Western” thoughts, this analysis draws inspiration from Melani Budianta’s (2019) view that, instead of “exteriorizing the West” in a tokenistic way, it is more meaningful to “(put) into dialogue whatever is usable from transnational critical theories with the local realities, and inter-referencing it with other Asian concepts” relevant to this analysis.[9] More than looking for some location in which to collectivize the contemporary individualist approach, kampong ethics addresses the urgent challenge of empowering sustainable trusted relationships within geo-spatial locations to foster collaborative applications of AI ethics formulations to in turn maximize AI governance

7 Eric Thompson (2004) challenges the “naïve” view that kampong life, as situated in so-called rural places, can be deemed as standing outside and apart from “modernity”, given that the “everyday social reality of its inhabitants is more akin to social life that is conceptually urban than not” (pp.2372-2373). See Thompson, Eric C. 2004. “*Rural Villages as Socially Urban Spaces in Malaysia*”. Urban Studies 41 (12). SAGE Publications Ltd: 2357–2376. doi:10.1080/00420980412331297573. The core theme of “time-space compression” on a local and global scale (Harvey, 1990 cited in Thompson, 2004, p. 2372) resonates with Homi Bhabha’s (1992) post-colonial theory of the liminal space— in a psychological and physical sense—marked by the recognition of the “world-in-the-home” and the “home-in-the-world” (Bhabha, Homi. 1992. “The World and the Home”. Social Text, no. 31/32. Duke University Press: 141–153. Available at: doi:10.2307/466222). Both sources reveal the porous nature of the boundary between local community life-spaces and the wider globalized world events.

8 Cheng, Chu-yuan. “*The Originality and Creativity of Sun Yat-Sen’s Doctrine and Its Relevancy to the Contemporary World.*” American Journal of Chinese Studies 10, no. 2 (2003): 149-62. Accessed May 19, 2021. http://www.jstor.org/stable/44289235.

9 Budianta, 2019, pp.252-253.

for social good.[10] To this end, this chapter approaches the local/global collaborative project of creating a harmonious AI-human life-space by drawing from whichever systems of belief prioritize the role of trusted relationships in stimulating communal modes of ethical decision-making. It is worth clarifying at the outset that rather than detailing the many and varied specific locations of AI in community to better understand our proposition for kampong ethics, which will be dependent on the nature of particular mutualities, purposes and receptions, this exploratory analysis restricts its applied focus to sites and dynamics of ethical decision-making about AI as a communal agent. Trust is vital as the fabric and outcome of these decisions.

> “More than looking for some location in which to collectivize the contemporary individualist approach, kampong ethics addresses the urgent challenge of empowering sustainable trusted relationships within geo-spatial locations to foster collaborative applications of AI ethics formulations to in turn maximize AI governance for social good.[10]”

Central to this analysis is the recognition that mutual trust can only manifest as an emergent quality of healthy social bonding in the community—both physical and digitally-based—as opposed to a fixed outcome or an end-product guaranteed by strict compliance to human control procedures injected into technology. In keeping with this aspiration, the following key contentions are addressed:

How Can the “Asian” Context Inform AI Principled Creation and Design?

For communality to represent an operational location while impacting the fundamental purpose of technology as a public good, AI creators should design/construct new technologies in service of the important communal relationships within specific communities.[11] For example, Singapore has consistently prioritized the value of respecting the elders in a social setting to a point where filial piety has become hard-wired into elder-care policy. In this scenario, it would be relevant for the design of AI technologies to maximize its accessibility functions for a smoother engagement with an older, perhaps less “tech-savvy” population, in appreciation of

10 Ibid.

11 No doubt such a transformation will face significant resistance not only from those who see AI as a market opportunity, but from the current AI tech hegemony happy to export their market dominance as some imperial, post-colonial mission under the guise of North/South world trading “freedom”.

the vulnerabilities and structural discriminators which can impede or at least colour the engagement between AI and elder populations. In keeping with such a collective conscience,[12] sensitive AI technologies need to address unique challenges faced by an ageing population (e.g. improve detection of common health issues, combatting loneliness and social isolation in old age, or even help the seniors better navigate and access services they require within increasingly digitalized or "smart" environments). For instance, the governments in Singapore, Japan and Korea had initially not predicted the pandemic's adverse impact on the elderly, the disruption to extended family support, and the challenges posed by smartphone track-and-trace technology for this demographic. So that the elderly were not discriminated against further in technology control agendas, it was necessary to improve the user-friendliness of the applications in order to better explain and accommodate the consequences of surveillance and social isolation.

To ensure AI actively participates in essential trust-building in the community, AI professionals cannot ignore imperatives to design/create new technologies in the service of social values and ethical principles that strengthen communal relationships. The question of which moral values can be deemed fundamentally "Asian" is a tricky one and likely to be reduced to tokenism if it becomes a required regulatory formula. It is more constructive to consider which ethical approach is most effective in stimulating healthy social bonding (from which shared trust emerge) within the varied contextual conditions and safe physical/digital spaces[13] of specific communities. By aligning their technologies with the ethical principles best suited to operate within the recipient community, AI creators and the tools they develop can better address the specific needs and priorities of who they are meant to benefit. Importantly, this approach steers away from the assumption that an abstract and universalist approach offormulating a single unified moral theory—often by a cultural outsider—can adequately describe the shifting ethical priorities of diverse communities in the local/global and physical/digital senses.

> "By aligning their technologies with the ethical principles best suited to operate within the recipient community, AI creators and the tools they develop can better address the specific needs and priorities of who they are meant to benefit."

12 This term is used as did Durkheim, discussed in Findlay M. (2017) Law's Regulatory Relevance: Power, property and market economies, Cheltenham: Edward Elgar. Available at: doi:10.4337/9781785364532.

13 The discussion of safe space and the conditions required for it in virtual and physical experience is complex. For the purposes of this paper, space is somewhere in which AI decisions are taken whether they relate to creation or application, and safety is reliant in part on ethical ascription.

What Can the "Asian" Mode Of Ethical Decision-Making In AI Regulation Be?

Culturally-aware AI and data decision-makers are required to recognize inclusive governance model(s) that best engage(s) the community it is meant to serve. In Singapore, the driving "Smart Nation" policy for urban design stipulates a citizen-centric focus for the deployment of AI technologies, and associated mass data sharing. Besides government policy, there are many factors which may directly or indirectly implicate the AI ethical attribution-distribution ecosystem, but integrating the concept of communality is essential to locating the ethical decision-making process, and thereby AI, within the community (characterized by its unique blend of socially-located contextual conditions). This approach embraces localized ethical and operational variance that contributes to the ultimate goal of ensuring AI technologies do not compromise, and instead, strengthen healthy social bonds of shared trust within the community. The presence of reasonable pluralism is not an outcome of regulatory failure in terms of a preferred model of community bonding, but a reflection of the diversity of thought and culture within the region. As such, values like openness, equality and compassion observed in a kampong context become fundamental to efforts in formulating a cohesive approach toward AI governance on a meso—and macro—level.

Where Does the Kampong Spirit of a Socially-Located Community Sit in Relation to a Global (Or Even a "De-Territorial" or "Socially De-Limited") Community?

Local communities are situated within a wider global community, but they are inter-operative via their cross-cultural influences. Since no social group recognizing the AI ecosystem exists in complete isolation, this analysis argues that applying the kampong spirit of a deeply felt and intuitive sense of commonality to local communities has a global relevance:

> Man must be envisaged not only in his relations with the State, but with the social groups of all sorts to which he belongs: family, tribe, city, profession, confession, and more broadly the global human community.[14]

14 Cassin, 1972 cited in Aroney, Nicholas. 2019. "*The Social Ontology of Human Dignity*". SSRN Electronic Journal. doi:10.2139/ssrn.3499573, p. 6.

This view of communality reconnects with Cotterrell's description of "social cement" as relying on complex, interconnected webs of trust which—motivated by its emotional foundations—will have a profound impact on AI/human interaction.[15]

Local/physical communities bleed into online/non-territorial communities because both entities are founded on social bonding which persists through time between people of similar interests, hobbies or values. As online environments call for a creatively different approach to regulate and minimize the impacts of AI technologies on people across various externalities, the kampong spirit of mutual support and solidarity rings true for such thinking. Decision-modes emphasizing common interest as a pathway to ease perceived (and actual) differences across certain geographical/political/religious/linguistic divides will be crucial in facilitating cross-communities collaboration on effective governance of emerging AI technologies.

Realistically, efforts emulating the kampong spirit by navigating the local and global community (which the modern human dwells within contemporaneously) may inevitably invite power dynamic issues, as they no doubt did in the purest kampong frame. In this scenario, the presence of distrust in different corners of community spaces plays a critical role in calling for key decision parties to address voices of dissent, often grounded in genuine concern for the potential negative social impacts provoked by technological change. Keeping true to the spirit of healthy and strong communal bonding, then, the decision-making process can manifest communitarian engagement where the lowest and the highest are called on to voice their opinions, even where established and often exclusionist hierarchies pre-exist.

Hegemonic Uni-Plastic Approach To AI And Data Ethics—North/South Imperialism—Market Elite Dominion

Where AI and big data are concerned,[16] the landscape of ethical guidelines and frameworks developed by the public and private sectors have been largely homogenous and often seemingly self-serving in a regulatory sense. Due to the power imbalance in international discourse on AI ethics, there has been underrepresentation and unequal participation from geographies such as Africa, South and Central America, and Central Asia.[17] These asymmetries can be seen as another feature of what has been deemed techno-colonialism from the AI-favoured nations and platforms to the world beyond. With the mainstream approach towards ethics more singularly embedded in what Polanyi referred to as "self-regulating markets" and their commodification/profit positioning,[18] rather than what John Rawls envisioned

15 Wong and Findlay, 2021.

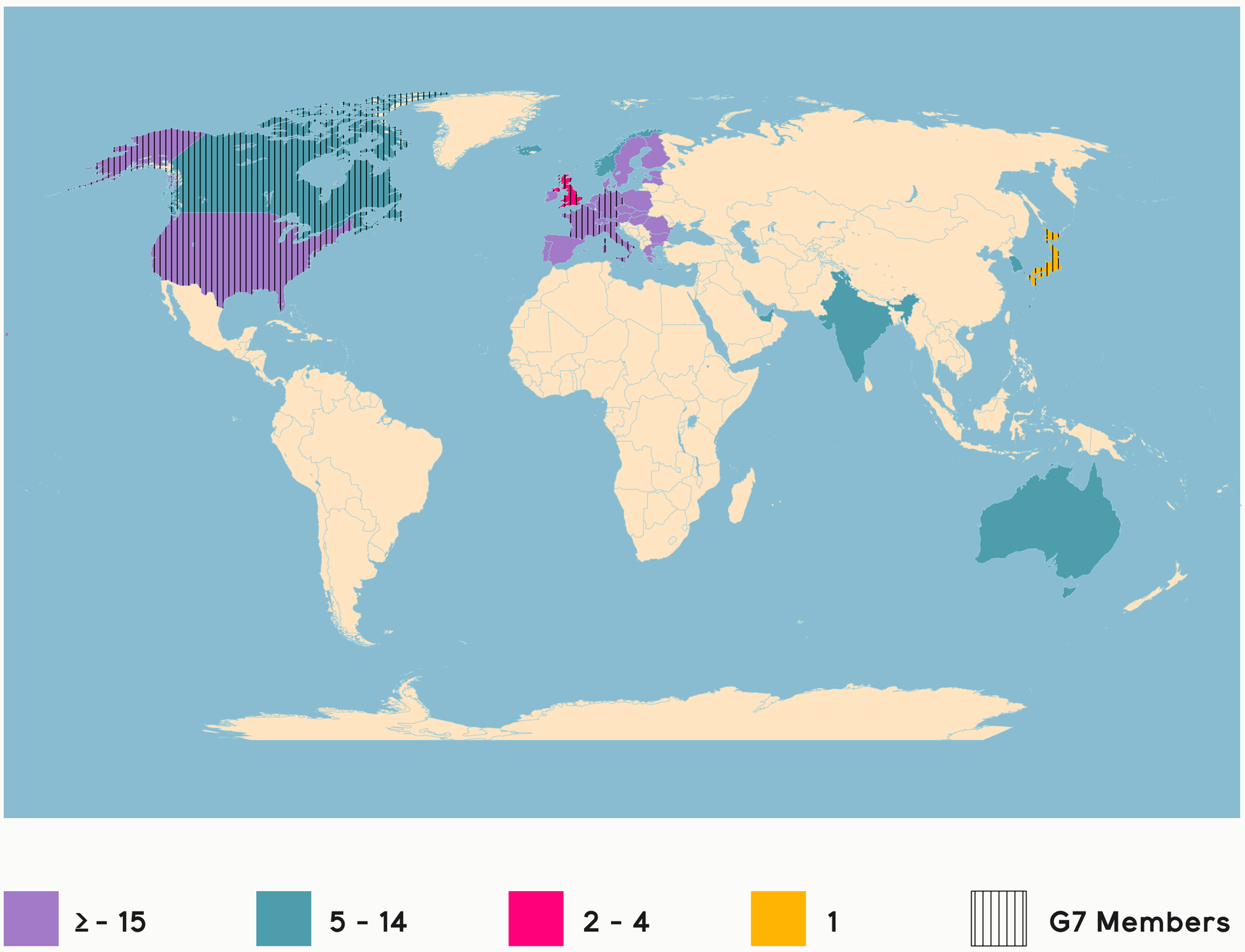
16 Again, AI and big data are used in a generic sense, as this analysis focuses on how ethics and trust influence AI creation and deployment, alongside the application of big data in terms of human choice. It is not the intricacies of the technology and data science, but the decision-making of their human agents against which our consideration of ethics and its communal positioning is directed.

17 Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "*The Global Landscape of AI Ethics Guidelines*." Nature Machine Intelligence 1 (9): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

18 Polanyi, Karl. 2001. The Great Transformation: The Political and Economic Origins of Our Time. 2nd Beacon Paperback ed. Boston, Mass: Beacon Press.

as “the fact of reasonable pluralism” (i.e., the irreconcilable religious, philosophical and moral doctrines which characterizes contemporary democratic and multicultural societies), the role of local knowledge and cultural pluralism has been missing in the global discourse.[19]

Figure 1. According to Jobin et. al. (2019), the geographic distribution of issuers of ethical AI guidelines by the numbers of documents released show most ethics guidelines are released in the United States, European Union, United Kingdom and Japan.



The sets of overarching ethical values and principles are largely similar, converging around the common themes of transparency, fairness, responsibility and trust. However, what is left unquestioned in the enterprise of dominant Silicon Valley-style iteration of AI ethics is the overarching spirit of instrumental reasoning.[20] In this position, technological inventions simply are a means to an end (i.e., as tools which enable human beings to influence, appropriate, and dominate the natural world).[21]

19 Audard, Catherine. 2006. John Rawls. London: Taylor & Francis Group. Available from: ProQuest Ebook Central.

20 Ihde, Don. 1979. Technics and Praxis. Boston Studies in the Philosophy of Science, 24. Dordrecht: Springer Netherlands. doi:10.1007/978-94-009-9900-8, pp.40-50, 66-81.

21 Merchant, Carolyn. 2014. “*Mining the Earth’s Womb*”. In Philosophy of Technology: The Technological Condition: An Anthology, edited by Robert C. Scharff and Val Dusek, 471–481. Hoboken: John Wiley & Sons.

But this ambition to use technological devices primarily as a servant of the neo-liberal market or as a tool for wealth creation remains,[22] standing in opposition to the ethical priority of achieving social cohesion through living in harmony with the natural environment which is held as a core value by many traditional communities in the world.[23]

The diversity in communal lifestyles across communities demands the communal location of AI-assisted technology (and the data on which it feeds) to be implemented with a more heterogeneous and elastic reasoning. If digitizing aligned with AI promotion is justified in communal terms, then both society and the economy need not compete to command the motivation for technology and its outputs. In such a setting, ethics goes beyond business compliance, or customer satisfaction understanding—it is a much more holistic determinant of trust in many inspirations, which sees AI as an active contributor to the social bonding processes in a community setting.

Comparatively, when viewed in the context of AI ethics with communal configurations, there are grave consequences in assuming that abstract ethical principles or a monolithic and universalist ethical framework can adequately encapsulate the lived experiences of diverse communities across the world. The assumption of "one true moral theory", which pays little attention to the context-dependent nature of ethical decision-making, may come at the neglect and sacrifice of existing plurality of reasonable ethical stances on any given issue already present in many communities across the world.

The capture of AI ethics by Western technological hegemonies represents another face in the techno-colonial advancement of AI from North to South worlds. It is problematic for AI ethics to become an accomplice in escalating the long histories of unequal power dynamics between "rich/poor" and "developed/developing" communities, wherein a small handful of political powers impose their constrained understanding of virtue and goodness upon communities deemed morally inferior through a racialized and othering lens. Polanyi would suggest that conceiving AI as embedded in the social presents a force for re-embedding other market relations and their now fictitious commodities back into socially relevant bonds (sustainable economic exchange rather than egoist wealth creation).

Contexts and Premise of Our Argument

In the analysis to follow, there is a divided stream of thought, interwoven and sometimes competing for prominence. In one flow, while cultural binaries are assuaged, there is an effort to identify features of decision-making at key points in the creation and deployment of AI, where the Asian "valuing" of community can

22 See chapter 6 of Findlay, Mark. 2021. Globalisation, Populism, Pandemics and the Law: The Anarchy and the Ecstasy. Cheltenham: Edward Elgar.

23 Saha, Arunoday. 1998. "*Technological Innovation and Western Values*". Technology in Society 20 (4): 499–520. doi:10.1016/S0160-791X(98)00030-X.

be highlighted to better ensure ethical attribution and distribution across the AI ecosystem.[24] Flowing alongside is the strong commitment that if communities can be determined by social bonds of trust, and ethical AI can stimulate trust, then it seems logical and compatible to locate AI within communities. This second stream reverts to the first when the arms link at an "Asian" understanding of community and its processes of decision-making to facilitate effective AI governance.

At times, the focus of what follows reads as if the inter-operation of social bonding leading to shared trust might be an "Asian consequence" of ethical decision-making. In fact, it is the other way around. Ethical decision-making can produce trust in communities and if the community consciousness is mutualized and inter-operative (as Asian communities can exhibit), then ethical AI is logically compatible with trusted community relationships.

The progress of this analysis is theoretical speculation without the benefit of ethnographic interrogation, and as such faces potential criticisms that the communal context for ethical decision-making is "Asian" only insofar as it seeks to offer an alternative approach to communal decision-making that is more compatible with our theory of community than conventional AI ethics and AI governance modes imported from elsewhere. In this view, our analysis may come across as suggesting (unintentionally) that communal decision-making can be considered "Asian" only insofar as it is not from "the West", thereby reconstituting the false dichotomy of East/West cultures that was previously rejected in the introduction.

> "Ethical decision-making can produce trust in communities and if the community consciousness is mutualized and inter-operative (as Asian communities can exhibit), then ethical AI is logically compatible with trusted community relationships."

If the premises and the argument concerning communal AI and ethical trust-building are tolerated, it may be for another paper to ground in more detail the social

24 Argued in detail in Orr, Will, and Jenny L. Davis. 2020. "*Attributions of Ethical Responsibility by Artificial Intelligence Practitioners*". Information, Communication & Society 23 (5). Routledge: 719–735. Available at: doi:10.1080/1369118X.2020.1713842.

meaning and value attributed to communal relationships in Asia, and thereby better extend our proposal that:

(1) AI professionals can/should integrate communality when designing and creating new technology for Asian communities, and

(2) the "Asian" mode of decision-making and governance is not one specific operation,

but whichever that effectively stimulates and strengthens the spirit and influence of communality within an Asian social environment. In addition, to give form to "AI in community", the locations of AI reception and the relationships emerging at the human/machine interface will require deeper ethnographies.

Our analysis begins with the proposal that communities are bonded through social relationships of mutual trust.[25] In their dynamic state, physical and digital communities are sustained by these social relations, which are also impacted by AI technologies and the big data on which they rely. For communities becoming more reliant on AI, it is vital to consider ways of formulating trust-generating frameworks which go beyond keeping "humans in the loop". Ethical principles said to govern the design and deployment of AI for communities (among other trust-generating frameworks such as morals, standards and even law) can be a measure of the extent to which the behaviours of AI systems and human agents lend themselves to generating shared trust in the community.

Recontextualizing the diverse manifestations of these social relationships of shared trust in the community to reflect the kampong spirit requires moving beyond superficial attempts at using "Asian" concepts as decorations to the fundamentally "Kantian" or otherwise Western ethical assumptions frequently hardwired into the dominant iterations of AI ethical frameworks. One example is using "Buddhist compassion" to compensate for the current weak spots in ethical attribution and distribution throughout the AI ecosystem,[26] which cannot be a philosophically robust exercise if the ethical frameworks in question still betray an overarching ambition for a technologically-facilitated global domination—a posture which runs counter to the core teachings of Buddhism.[27] Instead, this analysis proposes the adoption

25 Cotterrell, 2018, p.2.

26 Findlay, Mark, and Josephine Seah. 2020. "*An Ecosystem Approach to Ethical AI and Data Use: Experimental Reflections*" 2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G), 2020, pp. 192-197, Available at: https://arxiv.org/abs/2101.02008 or doi: 10.1109/AI4G50087.2020.9311069.

27 As discussed in Perry, Lucas, and Stephen Batchelor. 2021. "*Stephen Batchelor on Awakening, Embracing Existential Risk, and Secular Buddhism*". Accessed February 28. https://futureoflife.org/2020/10/15/stephen-batchelor-on-awakening-embracing-existential-risk-and-secular-buddhism/. In the effective account of Buddhist Compassion offered by Soraj Hongladarom, the solution to ethical AI requires a more holistic approach than using Buddhist values as a tokenistic label to decorate an otherwise un-ethical or non-compassionate machine (see Hongladarom, Soraj. 2020. "*AI for Social Good: Buddhist Compassion as a Solution*". Artificial Intelligence for Social Good. APRU. https://apru.org/wp-content/uploads/2020/09/layout_v3_web_page.pdf.)

of ethical decision-making processes which prioritize AI's communal embedding by implementing ethical principles that reflect the host community's vision, beliefs and values. By integrating communality into the creation and deployment of AI technologies, this proposed approach aims to maintain and strengthen shared trust in social relationships for the longevity and sustainability of the community.[28]

Kampong Social Bonding (Community)

Drawing from Cotterrell's analysis, the essential features of community are:

- **Demonstrating** some degree of endurance through time—social interactions may be limited in time and constantly fluctuating, but they are not merely fleeting and wholly transient.
- **Exhibiting** a degree of mutual interpersonal trust between the individuals involved in such relations (so that the relation imports a moral bond of some kind[29] between those involved in it).[30]

This definition of community is dynamic and inter-relational. Cotterrell's approach emphasizes the internalized and experiential dimension of social bonding in the community, which leans away from using spatially or temporarily constrained external identifiers—such as geography, nationality, ethnicity, culture, language, or religion—to define and identify a community. Accepting these factors can have profound ramifications on the formation of social bonds and interpersonal trust in a social environment, Cotterrell's interpretation of community becomes more open and receptive towards the social bonding between human-and-human(s) and human-and-machine(s) as distributed across the local/regional/global and offline/online dimensions of the "community".

Ethical Decision-Making Processes that Prioritize Communality

By conceiving of community as founded on social relationships, and not as a structure or mechanism alone for individualist well-being, this analysis proposes

28 The limits of this paper do not allow for a nuanced and satisfying examination of trust between AI and the human inhabitants of community, not even to the extent of distinguishing different forms and applications of AI into specific social needs. Sufficient for our present purposes is recognizing that what distinguishes communities are various external and internal forces, weak and strong bonds and organic or mechanical bonding agents at work on achieving (or undermining) the sustainability of trust.

29 While not having space in this analysis to do justice to the role of morals in social bonding, as we suggested earlier, it is important to recognize rather complementary frames such as morality when seeking to appreciate how ethics works in building trust.

30 Cotterell, 2018, p.2.

the conceptualization of AI ethics as encouraging decision-making that prioritizes, maintains and strengthens the preservation of the social bonds essential for the flourishing of individuals and of societies in the roll-out strategies of AI technologies. By aligning AI ethics to the shared norms valued by these social bonds, positive receptions for technologies that speak to the needs of community members become possible. Whether this translates to the adoption of certain moral values depends on the context, therefore taking an exclusive and prescriptive approach is counter-productive. What is most crucial in this picture is to recognize that communal displays of acceptance/disavowal, or trust/distrust of AI is a deliberate exercise in choice, returning to our decision-making focus. This presupposes a healthy and functioning society where citizens are free to participate in ground-level dialogues that can influence wider decision-making processes, but these conditions are not universal. Many communities operate under power hierarchies that are unconducive to these modes of ethical decision-making. As Amitai Etzioni clarifies:

> Communities, critics write, use their moral voice to oppress people, are authoritarian by nature, and pressure people to conform. However, from a communitarian point of view, informal social controls … ultimately leave the choice of violating social norms up to the individual, letting her determine whether or not she is willing to pay the social costs – as all innovators and social change leaders have – or conform. In contrast, state coercion pre-empts such a choice, as one sees in all oppressive regimes … [C]ommunitarians do not favor rolling back individual rights, but rather, paralleling them with concerns for the common good and the discharge of social responsibilities. To attain such a commitment, the values that are being fostered need to be truly accepted by the members and responsive to their underlying needs. If some members of the society are excluded from the moral dialogue, or are manipulated into abiding by the moral voice, or if their true needs are ignored, they will eventually react to the community's lack of responsiveness in an antisocial manner. In short, communities can

be distorted by those in power, but then their moral order will be diminished, and they will either have to become more responsive to their members' true needs or transform into some other non-communitarian social pattern".[31]

By contrast, there is room for individual and collective choices in communitarian societies channelling the village spirit in ethical decision-making. Indeed, pluralist communal ethics frames will not require some wholesale rejection of contemporary ethics incarnations or even their sometimes-problematic alliances with AI imperialism and neo-colonial technologizing. To require this radical revision a priori to the kampong spirit being achieved could well be reason for its rejection as utopian. While some ethical principles will be uniform and constant, appearing in many philosophical traditions even if expressed in different guises, their application to community conditions might be nuanced and reflexive, particularly as this analysis interprets ethics as an influence over exacting decision-making in identified decision sites.[32]

In allowing the communal location of ethics to motivate ethical decision-making, this alternative frame claims regulatory force through its ability to adapt socially relevant core values, meaning and beliefs to effectively stimulate shared trust through social bonding. Understood in this way, ethics can be perceived as core principled, communally dynamic and reflective of other contextual conditions that help engender trusted social bonds, such as morals, standards and law. In these terms, ethics washing cannot counter challenge community realities that undermine trust in AI, especially when the introduction of AI technology is inconsistent with other important contextual conditions that make possible the development of trusted social bonds in a communal setting, such as relationships of essential personhood or domestic spousal relations.

The Value of Reasonable Pluralism and Localized Ethical Variations

Cotterrell recognizes the process of social bonding as involving an affective dimension that inevitably combines convergent instrumental aims, shared beliefs and values, or merely a need for everyday coexistence and conformity with the

31 Etzioni, Amitai. 2015. "*Communitarianism*." In The Encyclopedia of Political Thought, edited by Michael T Gibbons. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118474396. See also: Etzioni, Amitai. 1996. "*The Responsive Community: A Communitarian Perspective*." American Sociological Review 61 (1): 1–11. https://doi.org/10.2307/2096403.

32 The decisions taken at different ecosystem locations to create and deploy AI and use big data for social good in community. Data science is already being discussed by philosophers of science in these terms: Leslie, David. 2021. "*The Arc of the Data Scientific Universe*". Harvard Data Science Review. Available at: doi:10.1162/99608f92.938a18d7.

customs or traditions of that environment.[33] For instance, a plethora of localized variations can emerge from each community developing their own shared and implicit understanding of the socially located values and meanings linked with certain objects/subjects and their actions/behaviours, alongside the emotions provoked by such context-specific connections. Trust generation and social bonding become diverse in manifestations and context-specific in how they influence human choices and decision-making.

A multi-variant and multi-dimensional view of community calls for a pragmatic way to fully capture the sheer diversity of lived experiences in the world when attempting to communicate, interpret and translate the mantra of "AI for social good" in regulatory discourses.[34] By understanding the need for AI ethics to authentically capture the global diversity of subjective engagement with AI technologies, the onus shifts to decision-makers in each community to engage in open, inclusive, informed and critical public deliberation on critical issues (such as human-tech integration) to help reconcile actual and perceived differences, provided if each community is committed to lean away from tribal ways of engaging in civic discourse.

A case in point is communities with greater capabilities to effectively govern and respond to the shifting dynamics of context-specific values, meanings and beliefs, in order to work these factors into optimal alignment with the adoption of trust-building AI ethics in important decision sites. The project of developing healthy, sustainable and resilient social bonding between AI and humans can become more meaningful, influential and legitimate by aligning ethical decision-making about AI (and the data it uses) with the specificities of local and regional communities. Thus, efforts in identifying and deliberating how AI technologies can better service their host communities can also improve public trust in the regulatory processes, if not in the specific technological entities themselves.

Without the commitment to locate AI in community, the roll-out of technologies for financial and political benefits can deeply undermine existing trust relations if the voices of the individual recipients and host communities are excluded in key decision sites and processes. Examples of elitist, exclusionist AI applications are myriad, from top-down digital transformation, to market-centred digital commerce. AI in community would require digital transformation to be owned by those it most affects, and the development of digital commerce to respect consumer interests and vulnerabilities. Trust requires ongoing maintenance because it arises from the choices of human recipients in the community to either accept or reject the deployment of AI technology and its purposes in the community; this runs counter to the dominant decision-making mode, wherein participation from data subject is missing and trust is seen as a product of top-down enforcement actions. As such, ethical AI decision-

33 2018, p.2.

34 Artificial Intelligence for Social Good. 2020. Association of Pacific Rim Universities (APRU). https://apru.org/resource/artificial-intelligence-for-social-good/.

making, as channelling and striving towards the kampong spirit of togetherness, requires the mutualizing of responsibility in using AI for social good.[35]

Virtual/Digital Communities

Similar to offline communities, social bonds of mutual trust between people are foundational in virtual communities. In a common transit, participants in virtual and physical communities share comparable desires for trusted social bonding. The social bonding process in safe digital spaces may be different to how they would arise through physical engagement, but it would be remiss in a pluralist version of communal location to ignore the pervasive significance of digital societies and virtual communities that span the North/South divide. With so much communal life being transacted in virtual realms, the influence of trust on social bonds is not diminished as a definitive characteristic of sustainable communities wherever located.

For Veronica Neri (2020), the virtual community is characterized by "dialectical relationships that constantly reshapes and redefines itself or the relational space" which introduces an added dimension of complexity (or instability) to the task of fostering healthy and sustainable forms of social bonding between individuals who can introduce themselves as they choose or replace their real identity with a different one at any given point in time.[36] It is important, therefore, to consider the porous intersection between the offline and online world to ensure trust-building efforts in the virtual community do not come at the sacrifice or even active destruction of healthy social bonding in offline communities.[37] This also goes to suggest that social ties in the intangible world such as the recognition of identity may take on similar presence but without the same reality.

Rather than trying to undo the increased entanglement between virtual communities and traditional offline communities, Neri suggests the right approach towards trust-building in a virtual community requires striking a balance in the individual and collective responsibility to resist against "(slipping) into indifference for oneself and for others, for the community and for technology in general".[38] This view resonates critically with the sense of solidarity in kampong ethics, where members belonging to the community—in online and offline settings—share a joint obligation to uphold social relationships of mutuality by recognizing trust as "the propeller of ethical growth".[39] Insofar as AI technologies can functionally reciprocate this quality of trust with its users—via AI safety and robustness monitoring—it can also be said that AI contributes towards the maintenance and improvement of shared trust within its recipient

35 Seah, Josephine and Findlay, Mark James, Communicating Ethics across the AI Ecosystem (July 29, 2021). SMU Centre for AI & Data Governance Research Paper No. 07/2021, Available at SSRN: https://ssrn.com/abstract=3895522 or http://dx.doi.org/10.2139/ssrn.3895522.

36 Neri, Veronica. 2020. 'Community and Trust in the Network Society. The Case of Virtual Communities'. In Trust, edited by Adriano Fabris, 137–150. Studies in Applied Philosophy, Epistemology and Rational Ethics. Cham: Springer International Publishing. Available at: doi:10.1007/978-3-030-44018-3_10, p.141.

37 Neri, 2020, pp.146-147.

38 Neri, 2020, p.148.

39 Ibid.

community. Thus, all agents (AI and humans alike) who contribute to communal social bonding have a direct stake in the status of shared trust in their virtual/offline community.

Conclusion: Kampong Spirit of Mutual Support and Solidarity To Shift AI And Big Data Into a Communal Resource[40]

A cynic would say the viability of communitarian location for AI (and the data it uses) is hindered by the neo-liberal global economic order as motivating both globalization and technological imperialism in the individualist terms of egoist wealth creation above all.[41] Such negative, anti-communitarian forces are not only seen in techno-colonial North/South terms in the hegemonies that drive technological imperialism, but are also alive and well in the elites of all cultures and market domains wherever located.[42]

Globalization has been captured by these exclusionist imperatives.[43] With the onset of each new global crisis to which AI solutions are directed, it becomes more pressing to confront and overcome the strain between the economic envisioning of the individualist and communal "self" via a reconceptualization of globalization (and AI's role across it) away from neo-liberal exclusionism and closer towards conscious shared humanity.[44]

Efforts in reconciling this deeply entrenched sense of separation—which stands in the way of authentic connection and puts a strain on existing social bonds in the community—demand a radical turning towards a "deeper collective truth of our interconnectedness".[45] In this view, the practice of self-inquiry is a necessary and indispensable part of debugging AI—without a long-term conviction of how humanity can and ought to transcend beyond its current, immediate struggles, AI will inevitably get caught up in the messy entanglements of conventional human biases, false opinions and contradictions.[46]

This challenge highlights the requirements for decision-makers to recognize ethical progress (for the purpose of fostering AI/human harmonious coexistence

40 Findlay 2017, chap 2.

41 Findlay 2021, chap 8.

42 Findlay, Mark, and Lim Si Wei. 2014. Regulatory Worlds: Cultural and Social Perspectives When North Meets South. Available at: https://ink.library.smu.edu.sg/sol_research/2670.

43 Findlay 2021, Chap 2.

44 Findlay 2021.

45 Perry, Lucas, and John Prendergast. 2021. "*John Prendergast on Non-Dual Awareness and Wisdom for the 21st Century*". Accessed April 19. https://futureoflife.org/2021/02/09/john-prendergast-on-non-dual-awareness-and-wisdom-for-the-21st-century/.

46 Ibid.

and creating AI for social good) as a shared goal for the individual and their wider community. That is, AI professionals, decision-makers and members of the public have the joint responsibility to cultivate their abilities to properly update personal beliefs, values, identity, and ethics in service of healthy forms of shared trust in the community (which correspond to what is true and good for all of humanity).[47] In doing so, it is possible to prevent the amplification of existing ethical and moral blind spots by increasingly powerful and prevalent technological systems and achieve a "regulative ideal of fully comprehensive, adequate emotional response" towards AI's role in the community.[48]

In keeping with Sun Yat Sen's notion of communal brotherhood/sisterhood and his visions of contextual equality and human welfare, it is not difficult to imagine, had he known of AI, that Sun would have sought out its place in a community that was neither strictly Eastern nor Western, but fundamentally human:

> Some Western scholars have contended that Sun's thoughts developed "on the frontier between China and the West". In their view, Sun's ideology appears to be a mere synthesis of Western ideas. This view touches only a part of Sun's idea. In an autobiographical sketch written in 1923, Sun described the formation of his own thoughts in the following statement:
>
> Among the various revolutionary ideas, I hold, some are adopted from traditional Chinese thought; others are appropriated from theories and practices developed in Europe, and still others are original insights grown out of my own critical reflections.[49]

Sun's Confucian adaptation holds that the state should not so much worry about poverty but unequal distribution; his conviction (from Kropotkin) that the progress of society results from mutual collaboration, as opposed to class conflict,[50] are examples of his polyglot philosophy which demonstrate the need to locate AI in communal relations of shared trust, respectful of pluralist ethical principles as not the

47 Ibid.

48 de Sousa, Ronald. 2001. "*Moral Emotions*". Ethical Theory and Moral Practice 4 (2): 109–126. Available at: doi:10.1023/A:1011434921610, p.124.

49 Cheng, 2003, p.150.

50 Cheng, 2003, p.153.

> That is, AI professionals, decision-makers and members of the public have the joint responsibility to cultivate their abilities to properly update personal beliefs, values, identity, and ethics in service of healthy forms of shared trust in the community (which correspond to what is true and good for all of humanity).[47]

province of North/South, East/West, but rather of the mutuality of human nature as crucial in social evolution (technological or otherwise).

With this in mind, we bring this chapter to a close with a short list of open-ended suggestions on AI ethics and governance approaches inspired by kampong ethics:

- **Recovering** the kampong spirit of solidarity allows policymakers to pursue ethical decision-making models working to overcome perceived divisions and seek common good, but this is possible only when key actors involved in the AI ecosystem see the value of trust and distrust as signals from the community. Furthermore, appreciations of AI ethics as influenced by trust should proceed from a specific recognition of the pluralistic nature of communities, where ethical priorities and the manner in which they create and sustain trust relationships will differ. As such, policymakers should work with this recognition and pursue empirically based enquiry focused on developing "AI in community" in their local contexts to counter the hegemonic forces of externally imposed standards by AI-favoured nations.

- **The** traditions of communities and neighbourhoods are more than ethnographic backdrops against which policy is developed. Instead, they provide an essential and historical understanding of social bonding processes. These communal contexts require the attention of policymakers striving for AI deployment approaches that are geared towards a genuine acceptance and empowered technological embedding into existing community relationships of trust. This may involve policymakers and AI promoters using online /offline communication pathways to address and resolve public sentiments of insecurity or concerns over AI roll-out into community spaces.

CHAPTER 07

Between Threat and Tool: The Poetics and Politics of AI Metaphors and Narratives in China

07

CHAPTER 07

Between Threat and Tool: The Poetics and Politics of AI Metaphors and Narratives in China

JENNIFER BOURNE AND
MAYA INDIRA GANESH

Abstract

Metaphors and narratives are world-building and future-envisioning modes of representation of artificial intelligence (AI), doing epistemic work by actively shaping the texture of reality. In this essay, we frame AI as 'poetically charged'; that is, its metaphors and narratives evoke philosophical and critical reflection on what it means to be human amidst machines designed to appear and act human-like. We bring this "poetic charge" to a study of metaphoric language and narratives of AI through select works of Chinese Science Fiction (SF) literature and in digital advertising and marketing campaigns. As we will discuss ahead, AI exists in past works of fiction that generate imaginaries of this technology before it was technically feasible; and contemporary "socio-technical imaginaries" continue to be generated as it evolves now. Hence, metaphors and narratives sourced from across time allow for a unique perspective on the shaping of AI in the socio-political and cultural context of contemporary China. We identify one pair of metaphors for discussion, a common one that exists in many different parts of the world: AI understood as both threat and tool. The popularity of this metaphor in China as elsewhere in the world suggests that governance actors critically review the positioning of China as substantially different from other countries; this may not be the case after all. We examine these metaphors in light of the human social conditions of life online in China, and specifically the conditions of digital labour. We propose that policy and governance actors critically assess the future of AI in terms of how the threat/tool dynamic refers to marginalized sections of Chinese society working as if they themselves were the tools of, and in, AI systems.

How Metaphors and Narratives Work

This essay is drawn from a larger research study that asked: How do technology histories, infrastructures, popular-scientific, and socio-technical imaginaries converge with local, metaphoric language about AI to shape future visions of this technology in different countries?[1] Through that study we argued for a richer, inter-disciplinary perspective on how AI is taking shape across 13 countries and nine languages. We found that metaphors are an entry point to understanding regional and cultural values, assumptions, and beliefs about AI. The study of popular metaphors and narratives of AI reveal a twinned dynamic at work: "poetics", which refers to how cultural, semiotic, and linguistic representations provoke reflection on what AI means for humanity and human society; and "politics" that refers to the political-economic factors containing and shaping the emergence of AI.[2]

> "The study of metaphors and narratives can sharpen our awareness of AI technologies as neither neutral, magical, nor inexplicable, and as entirely marked by social and local conditions shaping how this technology evolves."

1 This essay is drawn from a cultural research project conducted in 2020-21, the AI Metaphors Project, through a Berggruen Institute fellowship that both authors were involved with. We thank Nils Gilman of the Berggruen Institute for his insights in developing this work. This essay also builds on a set of workshops developed in partnership between the Global AI Narratives project at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge, UK, and the Berggruen Institute's China Centre at Peking University from Autumn 2020 to Winter 2021. These workshops brought together a small but distinct set of local and diaspora scholars, writers, and philosophers to discuss influences from Chinese Science Fiction literature and Chinese philosophy on the emergence and shaping of AI in China. More information is available online: https://www.berggruen.org/activity/ai-narratives-in-contemporary-chinese-science-fictions/

2 Brian Larkin's description of how the study of infrastructure brings poetics and politics together asserts that any infrastructure is always more than just technical functioning, and is also "concrete, semiotic and aesthetic", "with desire, fantasy" and "fetish-like aspects" encoded (p 335). Larkin, B. (2013) The politics and poetics of infrastructure. Annual review of anthropology, 42, pp 327-343.

In this essay, we assert that metaphoric language and popular narratives of science fiction are potent actors in constituting and shaping AI. Even though AI might be framed as a universal technology contained in the enduring cinematic imaginary of the humanoid robot, its opposite is actually true: AI is unfolding as multiple kinds of technologies, from natural language processing in digital assistants to app-based delivery work, across the world; robots might be entirely disembodied, or take non-human forms, and there are local ambitions and futures imagined for AI. In other words, technologies are not just neutral tools but are socio-technical,[3] and the study of language and narratives reveals how they wield power to build our desire for, and imaginations of, future worlds. Imaginaries are not deterministic plans for the future but they can influence social and policy ambitions in the form of "socio-technical imaginaries": "collectively held, institutionally stabilized, and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology".[4] A socio-technical imaginaries' approach identifies how scientific knowledge and technology innovations are quite intentionally made by different social, political, and economic actors through institutions, industry, culture, and knowledge-making practices. The study of metaphors and narratives can sharpen our awareness of AI technologies as neither neutral, magical, nor inexplicable, and as entirely marked by social and local conditions shaping how this technology evolves.

Metaphors, sometimes fragmentary and varied in form, refer to the use of a word or phrase for poetic, rhetorical embellishment, emphasis, or explanation. Metaphors describe, bring depth and emotion, elucidate contradictions, and identify tensions in a domain. Metaphors lead us to believe that the internet is a force for good because "information wants to be free"; or that "data is the new oil":[5] that cybersecurity is a matter of foreign invasion and bodily health[6] or burglary;[7] that social media is like a drug that we must struggle and deal with alone;[8] and that AI is both mysterious and threatening because it is a black box. Metaphors of fatherhood, midwifing, and reproduction were adopted by AI scientists and technologists of the mid-1980s

3 The "socio-technical" approach emerges from a body of theoretical work that argues that technology and society share a mutually co-constitutive relationship; so a technology artefact or process does not deterministically influence society, but is actually made through social practices, institutions, and relations, and in this process, society is re-shaped too.

4 Jasanoff, S. (2015) Future imperfect: Science, technology and the imaginations of modernity in Sheila Jasanoff and Sang-Hyun Kim (eds) Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power. Chicago and London: University of Chicago. pp 2-4

5 Puschmann, C., & Burgess, J. (2014) Big Data, Big Questions: Metaphors of Big Data, International Journal of Communication, 8(0), p 20.

6 Helmreich, S. (2000) Flexible Infections: Computer Viruses, Human Bodies, Nation-States, Evolutionary Capitalism, Science, Technology, & Human Values, Vol. 25 No. 4, Autumn 2000. pp 472-491

7 Wolff, J. (2014) Cybersecurity as Metaphor: Policy and Defense Implications of Computer Security Metaphors (March 31, 2014). TPRC Conference Paper, Available at SSRN: https://ssrn.com/abstract=2418638 or http://dx.doi.org/10.2139/ssrn.2418638

8 Sutton, T. (2017). Disconnect to reconnect: The food/technology metaphor in digital detoxing. First Monday, 22(6). https://doi.org/10.5210/fm.v22i6.7561

constructing AI as their child,[9] suggesting that it was something precious, like a demi-human to be nurtured into its full potential. Nick Bostrom's Superintelligence outlines AI as threat, clearly stating that there is no reason to assume that AI will be anything like humans, nor motivated by human desires and feelings. Making use of rich metaphors, Bostrom characterizes AI as a wise owl, and humans as hardworking, humble sparrows, who will one day find themselves paying for being in thrall of the owl's superior intellect. And that AI will regard human intelligence the way we humans perceive cockroaches.[10] Similarly, physicist Max Tegmark, founder of the Future of Life Institute, builds on Hans Moravec's original formulation[11] and describes AI as a "rising sea" of intelligence that threatens to swamp human ability.

We follow the work of German theorist, Hans Blumenberg, and US theorists, George Lakoff and Mark Johnson, who argue that metaphoric language structures our everyday reality, allowing us to conceive of new experiences, and take particular kinds of actions on that basis.[12,13] Metaphors intervene through structures of language by describing and communicating things we struggle to make sense of and for this reason, Science relies on metaphor to describe things that are unfamiliar, uncertain, or new.[14] But, metaphors can slip into common parlance so easily that we forget that they are indeed just metaphors; they can work as self-fulfilling prophecies. And it is near-impossible to come up with tests of metaphor accuracy, to see if they work or not, because that would imply being "outside" of the metaphor itself.[15] At best, metaphors suggest epistemic directions to follow in understanding how a technology works in society.

Narratives on the other hand are more elaborate and structured storytelling devices employing text, images, events, and different aspects of popular or niche culture. Narratives act as scaffolding for believable and desirable future scenarios that can influence human action, thought and social outcomes because of the power of storytelling. Storytelling can be a device to work through "difficult ethical questions, explor[es] anxieties, and experiment[s] with regulatory interventions in possible futures."[16] SF is a genre of literary and cinematic narratives that examines how scientific advances and technologies influence the emergence of future societies; AI is a distinct theme in SF from many parts of the world. Scholars find that in English

9 Curry Jansen, S. (1989) Mind Machines, Myth, Metaphor, and Scientific Imagination. Paper presented at the Annual Meeting of the International Communication Association, San Francisco, CA, May 25-29, 1989. https://files.eric.ed.gov/fulltext/ED311522.pdf

10 Bostrom, N. (2017) Superintelligence: Paths, dangers, strategies. Oxford, Oxford University.

11 View the diagram of the Rising Sea of AI here: https://www.researchgate.net/figure/Hans-Moravecs-illustration-of-the-rising-tide-of-the-AI-capacity-From-Max-Tegmark_fig4_330902196

12 Blumenberg, H. (1960 / 2010) Paradigms for a metaphorology. Robert Savage (transl). Ithaca, New York: Cornell University.

13 Lakoff, G. and Johnson, M. (1980/2003) Metaphors we live by. Chicago: University of Chicago Pp 124-130

14 Boyd, R. (1993) Metaphor and theory change: What is 'metaphor' a metaphor for? in Andrew Ortony (ed) Metaphor and Thought, 2nd edition, Cambridge: Cambridge University. pp 481-532.

15 We thank Nils Gilman for this insight. February 6, 2020.

16 Seger, E (2021) China workshop report. Global AI Narratives Project. https://www.ainarratives.com/resources/2021/1/13/global-ai-narratives-china-ii-2021

language SF, AI is constructed in terms of tensions and tradeoffs, such as between AI as a technology that might liberate humans from the burden of work, solve aging and disease leading to longer and healthier lives, and at the same time playing into a fear that humans will lose their humanity, become obsolete, and alienate people from each other.[17],[18] In this essay we discuss a similar tension and tradeoff in Chinese SF, of AI as both threat and tool.

Many popular narratives of AI - emerging from the North Atlantic, Western Europe, or Japan dominate the popular imagination, such as the Terminator movies, and Space Odyssey:2001, the robots R2D2 and C3PO in Star Wars, and the 1950s Japanese anime character, Astro Boy,[19] or the Ghost in the Shell manga series. They softly amplify local imaginations of AI-rich futures in terms of humanoid robots, some of them going rogue, and others being smart companions. Hence the leitmotif of AI as potential threat and/or tool are fairly widespread. These narratives provoke questions of where consciousness and intelligence reside, if the brain is a kind of computer that can be programmed, and could a computer then be brain-like?

> “Supposedly sterile or rational fields like policy, international relations, and science, have a distinct use of narrative to explore and assess future imaginaries, hopes, and fears.[22]”

SF narratives have gone from popular culture to being tools of state policy in the practices of security, geopolitics, and world-building. During the Cold War, the United States’ defence policymakers at think tanks like the RAND Corporation developed the “scenario”, inspired by SF narratives, as an instrument of strategic planning and decision-making.[20] This kind of planning is a hybrid of stories, role-playing, and war games; they “emphasise some ‘future history’” in “short narrative form” that organize[d] reality, and that allow “players” to “zero[ed] in on specific choices” and their outcomes, bringing texture to social realities in ways that “mathematical models

17 The Global AI Narratives Project has been documenting narratives emerging in fiction and non-fiction works from around the world, arguing that works of literature and fiction offer a view of how the future is likely to be shaped. See https://www.ainarratives.com/

18 The Royal Society (2018) Portrayals and perceptions of AI and Why They Matter. Report by Claire Craig, Stephen Cave, Kanta Dihal, Sarah Dillon, Jess Montgomery, Beth Singler, Lindsay Taylor at the Leverhulme Centre for the Future of Intelligence, http://lcfi.ac.uk/resources/portrayals-and-perceptions-ai-and-why-they-matter/

19 Sabanovic, S (2014) Inventing Japan’s ‘robotics culture’: The repeated assembly of science, technology, and culture in social robotics. Social Studies of Science, Vol. 44(3). pp 342-367.

20 Galison, P (2014). “The Future of Scenarios: State Science Fiction.” In Bolette Blaagaard and Iris van der Tuin (Eds) The Subject of Rosi Braidotti : Politics and Concepts, pp 38-46. London and New York: Bloomsbury Academic.

could not".[21] Thus, fictional narratives as a hypothetical sequence of events become instruments for practices of scenario-based planning. Supposedly sterile or rational fields like policy, international relations, and science, have a distinct use of narrative to explore and assess future imaginaries, hopes, and fears.[22]

Multiple, Local Futures

But futures are imagined differently everywhere. What is the AI-enabled future being imagined outside of technological superpowers like North America or Japan, such as in Nairobi, Buenos Aires, or Jakarta; or in our case, in China? Sometimes metaphors and narratives do not travel well because they are so closely tied to language and culture. For example, in China "Skynet" is the name given to an integrated system involving 30 million CCTV cameras used in law enforcement to apprehend criminals. However, in the Terminator series and franchise from Hollywood, "Skynet" is the name for a malevolent, all-seeing, all-knowing, synthetic super-intelligence.[23] While the Chinese Skynet is also putatively "all-seeing" thanks to its millions of CCTV cameras, it is framed as a net positive in keeping society safe.

AI has been portrayed as a "universal" machine that represents "universal" cognition. Ambitions of universality are evident in, for instance, Google Translate, a programme that has run into considerable problems precisely because it attempts to flatten differences between language structures assuming they are similar.[24] In identifying poetics and politics, we focus on AI's multiplicities in terms of something about the culture it emerges from. For example, India uses the metaphors of the lab and garage of the world to position itself as the place that will enable the global development of AI, based on its qualified success as a technology development centre.[25] "Lab" and "garage" suggest two very different perceptions of how science and technology get made: the former brings to mind a scientific research venue, controlled and well-managed; the latter, also part of Silicon Valley lore, is the messy space where creativity emerges so as long as genius innovators are left to just get on with it

Shazeda Ahmed notes that China's AI development is usually discussed by Western media in combative terms that convey anxieties associated with China's social

21 Galison, P. (2015) "The Half-Life of Story." In Tessa Giblin (Ed) Hall of Half-Life, Graz: Steirischer Herbst, p 99

22 For instance, see Carol Cohn's 1988 ethnographic account of "Sex and Death in the Rational World of Defence Intellectuals". In this paper, she takes a critical feminist approach to the role of language in the work of "defence intellectuals", the men who created the theory that "informs and legitimates America's nuclear strategic practice" (p1). She discusses the unsettling use of metaphoric language, such as in how planetary scale destruction is understood as "clean", or in the strongly heteronormative sexuality: https://escholarship.org/uc/item/83k4763m

23 We thank Shazeda Ahmed for this insight. February 6, 2020.

24 Christoph Ernst, Jens Schröter and Andreas Sudmann (2019) AI and the Imagination to Overcome Difference. Spectres of AI, #5 Spheres: Journal for Digital Cultures https://spheres-journal.org/contribution/ai-and-the-imagination-to-overcome-difference/

25 Expert research interview with Dr. Urvashi Aneja. October 22, 2020.

credit system, widespread application of facial recognition technologies, or automated censorship of speech online.[26] There are strong metaphors of hygiene in policy discussions in the US about China in the context of AI; for example, that of "clean" networks "free" of Chinese technology.[27] China, the US, and Europe are also already positioned within the narrative of an "arms race" that only amplifies existing nationalisms and precarious geopolitical tensions.[28] However, these accounts in international media must also be understood as local and partial and built on centuries of prejudiced stereotypes, violence, and clash-of-civilizations style arguments.[29]

However, Chinese cultural engagements with AI are less visible as a site of discussion about this technology; so, in this essay we present insights from having reviewed 36 Chinese SF novels published between 1986 and 2019 that were either nominated for or won China's Galaxy Award.[30] We also refer to three recent popular SF stories, a short SF collection,[31] and digital materials from the tech industry and popular culture. Across this data, we find one set of metaphors emerging, that of AI as evoking uncertainty and tension because it presents as both threat and tool. We analyse this in light of recent research about the conditions of app-based delivery work. To reiterate, while metaphors might reveal something about how we imagine AI, this technology is already emerging in a socio-political-economic context, and this becomes an opportunity to reflect back on the metaphor itself and how it is shaping reality.

Threat and Tool in Chinese SF

A little like the English language metaphor of AI as a "black box", which suggests something inscrutable or ominous, in Chinese, AI is likened to the Greek myth of "Pandora's Box" (魔盒), which is itself a metaphor for something that could be dangerous, but that also eventually contains glimmers of hope. This foreshadows the notion that AI is something that must be managed or constrained so as to limit and manage the threat it presents. But what is the threat and how it is being managed? The threat thought to be associated with AI is that what starts off as

26 Ahmed, S. (2019) The messy truth about social credit. Logic, Issue 7/ China. https://logicmag.io/china/the-messy-truth-about-social-credit/

27 Whittaker, M., Ahmed, S., Kak, A. (2021) China in global tech discourse. AI Now Institute. https://medium.com/@AINowInstitute/china-in-global-tech-discourse-2524017ca856

28 Cave, S., & ÓhÉigeartaigh, S. (2018). An AI Race for Strategic Advantage: Rhetoric and Risks. 36–40. https://doi.org/10.1145/3278721.3278780 Also see Kaltheuner, F. (2021) A new tech cold war? Not for Europe. AI Now Institute. https://medium.com/@AINowInstitute/a-new-tech-cold-war-not-for-europe-4d4f2f8079b6

29 Mackereth, K (2021) Nationalism., In AI Now Institute (Eds) The New AI Lexicon https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-ai-nationalism-417a26d212f8

30 Of 98 novels nominated for the Galaxy Awards, eight novels could not be located as digital versions; and 54 of them did not relate specifically to AI.

31 Folding Beijing, Waste Tide and Eternal Hospital are recent, popular SF stories available online. The 2021 short story collection, AI 2041, by Kai-fu Lee and Chen Qiufan, was also reviewed.

mimicking the human will one day supersede it, and without the opportunity for appeal. The narrative of threatening robots is offset by the theme of control. Thus, a robot worker/helper that is engineered to work for humans must be kept in check to maintain social harmony and order; this is perhaps a thinly-veiled acquiescence to the ruling party's consistent and overt control over society.

Fears associated with the threat of AI in the form of humanoid robots going rogue are evident in Computer Monster King (电脑魔王) by He Hongwei (何宏伟) (1993). In this, the protagonist Qiu Xian lives in the M79 zone on earth where life is controlled by a computer system called Beth. He is assigned an attractive robot wife who follows his orders and does what he pleases. After a series of unexpected incidents, Qiu's robotic wife is killed and he wanders into the M80 zone where people are not controlled by Beth. After some initial adjustments, he finds that he enjoys life at M80 more. Beth tracks him down, destroys M80, and kills most residents on it. Qiu eventually discovers a flaw in Beth and destroys it.

Or, consider, The Way of Machines (机器之道) by Jiang Bo (江波) (2015), which imagines a future where humans discover how to replicate cognition and build immortal bodies, thus becoming "robotmen" (机器人). Before long, the "robotmen" start killing each other. The few last humans who were not "robot-ised" escaped to a virtual world. They eventually were able to combine their minds together into a "brain pool" and defeat the "robotmen". Space Monastery (太空修道院) by Tan Li (谭力) (1991) is about Didi and Jiejie, two obedient robots that guard a monastery in space. We could speculate that this novel being awarded the Galaxy Award in 1991, just two years after the government crackdown on the 1989 Tiananmen Square resistance, is an acknowledgment of the values of obedience to the ruling party. Similarly, in Superbrain (超脑) by Zhao Ruhan (赵如汉) (1995), scientists are teaching the eponymous computer system emotions and cognitions. This system takes up hundreds of acres of land in Tibet. But a group of "naturalists" here are concerned that the human race may be controlled by Superbrain in the future. But Superbrain proves itself to be loyal to its creator and merges with its creator to become a superpower ("God of Computer") and defeat an alien invasion. Again, the narrative of a robot/AI takeover ends with the preservation of human power. More importantly, the doubts of the Tibetan "naturalists" are quelled.

Like narratives in many parts of the world, these popular tales position non-human super-intelligence as threatening to humans, and eventually centre the triumph of human intelligence. But these stories exist alongside metaphors of AI emerging in business advertising and marketing as a tool that will work for us. For instance, Microsoft advertises AI as a tool that comes into its full potential when picked up and used by humans. The tagline of their advertising spot is, "What's a hammer without someone to swing it?".[32] Here, AI is activated by human skill and creativity,

32 The advertisement can be viewed on Youtube: https://youtu.be/jsyWtLjmfK8 We reference the work of Dr. Gladys Pak Lei Chong, at Hong Kong Baptist University, who made these points in presenting her research at the Histories of AI: A Genealogy of Power, the Mellon-Sawyer seminar series at the University of Cambridge, https://www.hps.cam.ac.uk/research/projects/histories-of-ai November 20, 2020

and thus becomes a tool that is enhanced by, and enhances, the human. Critically, this positions AI as a neutral tool, something that has potential when and if activated correctly. (But this also suggests that incorrect activation can result in negative outcomes.) In the Microsoft advertisement, we never see the kinds of images commonly associated with the visual culture of AI, such as glowing digital brains or robots; we see just everyday people. However, we see the kinds of humans who are the intended customers of Microsoft's products, like artists, business managers, and teachers. We do not see the humans who actually constitute the labour force that enables AI, like the online workers who tag and label images for computer vision, test drivers of autonomous vehicles,[33] platform-based workers, and factory workers.[34] As we will discuss ahead, the people who constitute AI's workforce are actually the tools, so to speak, inside AI. Thus, the advertising of AI as a tool elides the reality of AI as infrastructure that is constituted by labour, and emphasizes imaginaries of productivity.

The AI-as-a-tool metaphor is framed as an upgrade to everyday life, as "AI plus"; the plus refers to all the benefits it will bring.[35] So it is not surprising that Chinese businesses position AI as a "friendly partner that will help us get our work done".[36] In popular media, AI is likened to something exciting, like the feeling before the wedding night, or the feeling of expectant parents who look forward to that which will accompany them in the future[37] （像新婚前的忙碌与忐忑, 像出婴儿出生前的父母心情, 未来却是相随相伴. AI is positioned as a "work assistant" and "life assistant" that is full of love[38]（既是工作助手, 也是生活帮手—— 这样的人工智能更有爱). But these cheerful business media metaphors elide the long history of gendered labour being replicated by female voice assistants. Giving orders to the women who do menial, basic tasks around the house to care for us, from domestic help to our mothers, is a common experience that is unquestioningly replicated here.[39] For instance, in one of many advertisements for the digital assistant, Xiao Ai, by Xiaomi, the assistant is female-voiced, deferential, and follows orders delivered by the mother of the house, not unlike a domestic worker that middle-class Chinese might have in real life. In another advertisement targeted at a young, wealthy, and urban demographic, Xiao Ai is clever, helpful, and has a dry sense of humour. After the human flops into bed at night and amuses herself by

33 Ganesh, MI (2020) The ironies of autonomy. Nature Humanit Soc Sci Commun 7, 157 (2020). https://doi.org/10.1057/s41599-020-00646-0

34 Delfanti, A. & Frey, B. (2020). Humanly extended automation or the future of work seen through Amazon patents. Science, Technology, & Human Values. pp 1-28, https://10.1177/0162243920943665

35 Online interview with Dr. Julie Yujie Chen, February 3, 2020

36 Email interview with Jennifer Pan, January 2020.

37 Laoji890. (2021, September 10). 人工智能（AI）这个让人恐怖的名字，仅仅就是一个美丽的比喻吗，是人们自扰的? 知乎. https://www.zhihu.com/question/270841439/answer/357176185

38 Sheng, S.T., Yi, S.R., & Xu, J. (2021, August 15). 既是工作助手，也是生活帮手—— 这样的人工智能更有爱. 新华网. http://www.xinhuanet.com/tech/

39 Lingel J., and Crawford, K. (2020) "Alexa, Tell Me about Your Mother": The History of the Secretary and the End of Secrecy. Catalyst Journal. Vol. 6 No. 1 (2020): Special Section on Chemical Entanglements: Gender and Exposure https://doi.org/10.28968/cftt.v6i1.29949

ordering Xiao Ai to turn the lights on and off many times, the digital assistant pipes up to say it will even count sheep for her if she cannot sleep.[40]

> “The advertising of AI as a tool elides the reality of AI as infrastructure that is constituted by labour, and emphasizes imaginaries of productivity.”

Poetics and Politics of Threats and Tools

The metaphor of AI as a tool emerges in another slightly different socio-political context that has significance for our discussion. A report about the Chinese government’s use of automated censorship tools to regulate online discussions carried an intriguing metaphor: that while the human is a machete, AI is a scalpel.[41] This refers to the precision and efficiency of algorithmic keyword censorship over that of human-operated digital censorship. But there is an interesting angle to this that brings into focus something beyond AI metaphors as offering simple comparison between human and machine intelligence in terms of speed or skill. There has historically been a flowering of Chinese civil society responses to censorship that leverage the intrinsically poetic structures of Chinese language, in which homophonetic words can convey different meanings. In 2009, Chinese netizens created a shadow lexicon, the Grass Mud Horse lexicon, that ridiculed government censorship and became a wildly popular local cultural phenomenon.[42] “Grass-mud horse” (cáo nǐ mā) is a goofy mythical creature whose name sounds nearly the same in Chinese as a specific insult.[43] More recently, Xiaowei Wang writes that when the hashtag #MeToo was algorithmically censored in China, people used the combination of the rice bowl and rabbit emojis to get past the censors. Rice in Chinese is pronounced as “mi” while rabbit is pronounced as “tu”. When read aloud, the combinations sound like “me too”.[44] In other words, human practices of evading censorship rely on wit, subtlety, and creativity. This is a unique kind of in-group language that originates spontaneously and does not have to be explicitly learned; however, a natural language AI system would have to be explicitly trained to

40 https://youtu.be/ATat12_gRtk

41 Cadell, C (2019) China’s robot censors crank up as Tiananmen anniversary nears. Reuters, May 26, 2019, https://www.reuters.com/article/china-tiananmen-censorship-idUKL4N22W11L

42 See: https://chinadigitaltimes.net/space/About_the_Grass-Mud_Horse_Lexicon

43 Wines, M (2009) A dirty pun tweaks China’s online censors. The New York Times, March 11, 2009, https://www.nytimes.com/2009/03/12/world/asia/12beast.html

44 Wang, X (2019) The future of memory in Rigged Systems, a web residency at Akademie Schloss Solitude https://www.akademie-solitude.de/de/formats/the-future-of-memory/

identify these neologisms, which would take far too long and negate the very point of automating censorship.

The threat/tool understanding of future AI also provokes us to go deeper into examining how they play out in current social relations and arrangements. Hao Jingfang (郝景芳) is a well-known Chinese SF author who won the prestigious American Science Fiction Prize, the Hugo, in 2016, for her short story, Folding Beijing (北京折叠). This thinly-veiled discussion of socio-economic inequalities in contemporary China resonates with the current and ongoing struggles of gig and platform-based delivery workers who, studies show, are treated as tools. And, like the netizens resisting keyword censorship, they activate unique tactics of resistance to AI's actual—rather than threatened—harms.

Folding Beijing is set in a Beijing that is constituted by three sub-cities that occupy the same physical space at different points in time. From 6am on one day till 6am the next, five million residents of "First Space" live in a luxurious and elegant Beijing with wide, gingko tree-lined boulevards and palatial homes. Then, First Space folds up and gets swallowed into the earth while its residents fall into a deep sleep in their specially designed sleep pods and under the influence of a soporific drug. In its place, "Second Space" pushes up and unfolds, a parallel city of 25 million people who occupy the 6am to 10pm slot. Second Space is less wealthy than First Space, but is still comfortable and clean. Then, Second Space falls asleep and folds in, and "Third Space" emerges, a city of 50 million people working from 10pm to 6am. The residents of Third Space occupy the graveyard shift because their work is to clean up after and maintain First and Second Space. Third Space is for the underclass; its citizens live cheek-by-jowl, are discriminated against and cannot imagine ever saving up enough to escape their reality. There is also nowhere to escape to. The reason Beijing can exist as three cities is because of the impenetrable borders and maintenance of social differences between them. But, Folding Beijing's protagonist Lao Dao, risks life and limb trespassing into First Space for the first time in his life. Lao is so overwhelmed by the stark differences that his experience in First Space is dreamlike, disorienting, and confusing. He gets a peek into how the elites of First Space make decisions about the tens of millions of futures in Third Space, chiefly to secure their own power and privilege. At one point in the narrative, Lao overhears First Space elites discussing the implications of automation for jobs and employment (in Third Space). They rehearse various arguments about the macro-economics of the balance between GDP, social welfare, and automation, but in terms of completely botched data analytics. The elites discover their mistakes eventually, but there remains a lurking doubt: how many errors went undiscovered, and who lives with the costs?

This fictional narrative of some humans living under conditions of inequality and servitude is not so fictional considering a growing body of research and public discussion about the plight of delivery workers employed by the gig, platform, or "on-demand" economy in China. The metaphor of AI as a tool to take on tasks for humans sits awkwardly alongside the increased algorithmic control of humans, who feel that they themselves are "automated" and are treated like mere cogs in the

machine. Research finds that delivery workers' time, bodies, work, and wages are algorithmically managed and regulated.[45] Similarly, "Delivery Riders, Trapped in the System" (外卖骑手, 困在系统里) is an anonymous and collectively-authored report in the news magazine Renwu (人物) that documents the "horrors" of food delivery workers trapped in a system they cannot control: taking risks in traffic to meet targets; being physically injured or exhausted by work; feeling out of control by the gamified logics of algorithmic targets that promise a marginal increase in pay for much more work, and all this without secure or fixed employment contracts.[46] The pressures of delivery work come from a simple, punishing logic: complete as many deliveries as possible within the shortest period of time. Eventually, these platforms are hardly neutral tools and are architected to prioritize profits for businesses that own and develop them. But the logic of working for these platforms is organized around the black-boxed dictates of the app, as if the app were "neutral". But, delivery workers adopt tactics to gamify the system to circumvent algorithmic control by physically protesting, tactically supporting each other to complete orders on time, and working on multiple platforms to track and pick up the most feasible delivery work.[47] Thus, not unlike citizens who organize to avoid algorithmic censors, workers similarly organize to evade algorithmic bosses.

The realities of labour in AI contexts illuminate the narrative of threat and/or tool. We could suggest that narratives anxious about robot uprisings are a reflection of concerns of management but are framed in terms of an existential threat to humanity. Is the threat of workers—treated as tools—rising up to resist? Early Western cinema about AI, notably the 1927 German film, Metropolis, and Karl Čapek's stage play, R.U.R, which gave us the word "robot" (from the Czech, Robota, meaning slave), embodies the concerns of workers uniting to resist oppression. For, as studies of automation indicate, human workers have for decades been increasingly displaced—rather than replaced—with waves of automation. In the Euro-American context, this has been a narrative about robots taking away human jobs; but as automation theorist Aaron Benanav argues, the jobs are just not there to be taken away. The problem is not technological change, but a decades-long marinade of economic stagnation, austerity, defunding of social and public infrastructure, and debt, among others.[48] "Robots taking away jobs" could just be a metaphor for states' anxieties about their own identity and resources in a changed world. If automation increases, what is the responsibility of states to displaced workers?

A slightly different future of work is imagined in The Job Saver, a short story in the collection AI2041 by futurist Kai-fu Lee and SF author, Chen Qiufan. The authors imagine that AI will "decimate" routine jobs necessitating a new industry: job

45 Sun, P. (2019). Your order, their labor: An exploration of algorithms and laboring on food delivery platforms in China. Chinese Journal of Communication, 12(3), 308–323. https://doi.org/10.1080/17544750.2019.1583676

46 We have referred to the translated version of the Renwu long form investigation published in the English language blog, Chuang, https://chuangcn.org/2020/11/delivery-renwu-translation/

47 Op Cit, Sun P (2019)

48 Benanav, A. (2020) A world without work? Dissent Fall 2020. https://www.dissentmagazine.org/article/a-world-without-work

reallocation firms. These firms will help retrain and reassign displaced workers. However, it is possible that those reassigned jobs will not be satisfying to humans' desire to feel productive and useful, and eventually, even those menial jobs will disappear as well. So, a video game is created to fool workers into believing they are doing something more valuable, such as construction work in "third world countries". But in reality, the game is designed to consume workers' time and make them feel like productive members of society. This story assumes that AI and automation will continue to develop with negative outcomes for the human worker.

Conclusion

Scholars note that Chinese development and application of AI technologies have generated antagonistic concerns associated with its geopolitical, neo-colonial, and business influence; Kerry Mackereth identifies a distinctly Orientalist and racist register to the framing of these concerns.[49] Somewhat motivated by this, we undertake a different approach here, turning to the country's own cultural materials and social contexts of technology use to understand the nuances associated with the emergence of AI. Reading 36 award-winning popular works of SF published between 1986 and 2019, short stories, and business media such as digital advertising, we identify a dominant construction of AI as potential threat, or tool, a dynamic that is evidenced in dominant Western narratives of AI as well.

The study of metaphors and narratives allows us to momentarily and critically assess the realities we are trying to understand and navigate through our use of metaphoric language. This becomes our methodological approach in this work to examine the "socio-technical imaginaries" emerging in the connections between the interior, poetic, associations we have of technology as conveyed through language and culture, as well as its broader social and political-economic dimensions. Science Fiction is a rich source of narratives that can both inspire and threaten but are not necessarily blueprints for the future. Paying attention to metaphoric language offers evidence of the beliefs we have about the world and our actions in it. If we consider AI to be "summoning a demon",[50] that it will swamp human abilities, or take away jobs, then it is most likely we will be defensive, and work to mitigate such threats. The notion that AI is a future threat has been shaped, primarily, by those who are ignoring the reality that AI is already harmful to communities, such as delivery workers. Metaphors and narratives are not free from what Sandra Harding refers to as "social fingerprints", that is, a trace of where metaphoric thinking comes from, which then defines what is considered to be true about AI.[51] So, whoever sets the terms of the language or

49 Mackereth, K (2021) 'Nationalism' In AI Now Institute (Eds) The New AI Lexicon https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-ai-nationalism-417a26d212f8

50 The Silicon Valley tech mogul, Elon Musk, said in 2014 that he considered AI to be "worse than nukes" and like summoning a demon. See: https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/

51 Harding, S. G. (1986) The science question in feminism. Ithaca: Cornell University.

amplifies particular narratives is likely to shape agendas for science and technology innovation, development, and policy. In China as in other parts of the world, there is an enthusiastic business narrative that promises a better life with AI. Reading SF literature gives us insight into the interior condition experienced by humans enmeshed in and with the future of automation. So, what might delivery workers, who constitute the assemblages of "tools" in AI, generate as metaphors of AI, if asked? Further policy work could engage communities of gig and kinds of digital workers to continue to document and discuss their experiences of work, and how else to shape AI futures from their perspective and concerns.

We could conclude that while similar business dynamics might actively shape the emergence of AI in many parts of the world, it is the particularities of history, culture, and language that suggest how to attend to those most vulnerable to the harms of automation. Eventually, cultural narratives and metaphors draw attention to the human condition and inspire reflection on what it means to be subjected to the visions of powerful metaphor-makers. How this inspires policy and governance to action must be a matter of our politics.

> "The notion that AI is a future threat has been shaped, primarily, by those who are ignoring the reality that AI is already harmful to communities, such as delivery workers."

Digital Futures Lab is an interdisciplinary research collective that interrogates the complex interaction between technology and society in the global south. Through evidence based research, participatory foresight, and public engagement, we identify pathways toward equitable, safe and caring futures.

Instagram Twitter Linkedin **Website**

The Konrad-Adenauer-Stiftung (KAS) is a political foundation of the Federal Republic of Germany, which has, for over 50 years, committed itself to the promotion of democracy and international cooperation. The Rule of Law Programme Asia, based in Singapore, is dedicated to working with its Asian partners toward the development of rule of law in the region. One of the particular areas of focus is to explore the interplay between technology, society and the role of law.

Facebook Twitter Linkedin **Website**