Association of Pacific Rim Universities (APRU)

AI for Everyone Project

Benefitting from and building trust in the technology

Project Overview and Policy Statement[1]

November 30, 2018

Cochairs: Toby Walsh, University of New South Wales
Jiro Kokuryo, Keio University
Project Director: Catharina Maracke, Keio University

[1] This document has been created as a result of a discussion series supported by Google. Statements and opinions expressed are those of the authors of this document and under their full responsibility.

1 Project Overview

- If governed adequately, AI (artificial intelligence) has the potential to benefit humankind enormously. However, if mismanaged, it also has the potential to harm humanity catastrophically. Drawing upon this realization, the Association of Pacific Rim Universities, with the support of Google, launched an international collaborative research project in 2017.

- The project started with a call for papers, which resulted in the submission of twelve working papers (see Table 1 for the list of authors and titles). Two drafts were made for each paper, which received feedback from other members ("coaches"). Completion of these had been scheduled for the end of November 2018. The present report is based on the second draft, submitted for the brush-up session held at Hong Kong University of Science and Technology on September 1, 2018.

- The first meeting of all contributors was held at Keio University on December 1st 2017. Over the course of the discussions, all contributors agreed that the title of the project should be "AI for everyone: benefitting from and building trust in the technology". This title reflects the belief that "access to the benefits of AI, awareness about the nature of the technology, governance of the technology and its development process with a focus on responsible development, should be transparent, open, understood by and accessible to all people regardless of their geographic, generational, economic, cultural and/or other social background."

- Four sources of societal threats that need to be addressed were identified to which analyses and suggestions were offered by the academics.
  1 "Black box" machines manipulating human society
    Technical solutions to enable humans to understand, explain, trust and control the behavior of AI, reducing their perception as " black boxes" were offered. As well, a framework for the development of a certification system for trustworthy AI systems was proposed.
  2 Unethical uses of AI
    Moral, technical and legal solutions to the malicious unethical use of AI were discussed. The importance of responsibility of "moral agents" for the use of AI in war and political manipulation was explored. A proposal was offered to update the notion of legal culpability.
  3 Renewed threats on privacy
    The dangers of inference attacks in identifying and learning patterns in data to predict attributes of subjects or databases were identified. Suggested technical solutions were made including ways to integrate data scattered across various organizations while protecting the privacy of individuals. A philosophical and historical overview of the tension between individual and collective benefits of managing data was provided.

4 AI may widen the gap between the rich and the poor
The prospect of the "replacement of human beings" by AI was discussed along with a strategy to educate and reeducate the new and existing workforces to equip them to deal with the emerging needs.

- While many of the issues and analyses are global and universal in scope, some distinctly regional characteristics have emerged throughout the discussion process. A particularly important divide seems to exist in attitudes toward individualism and autonomy, which are perceived and institutionalized in various ways, giving rise to different attitudes toward the use of technology in state surveillance for public order. In response, a policy statement calling for sensitivity to political and cultural diversity has been put together (see Appendix 1).

Table 1: List of Working Papers

Author and Affiliation	Paper Contributed	Coaches to the Paper
Dremliuga, Roman Far Eastern Federal University	How Development of Artificial Intelligence Technology Will Cause Changes in Crime and Criminal Law	Erskine, Monroy, Lau
Erskine, Toni The Australian National University	Flesh-and-Blood, Corporate, Robotic? Moral Agents of Restraint and the Problem of Misplaced Responsibility in War	Shen, Yap, Dremliuga
Gal, Danit Peking University/Keio University	Best Bot Friend (BBF): The Emotional and Social Implications of Socializing with an AI	Yang, Shokri, Monroy
Lau, Chong-Fuk The Chinese University of Hong Kong	The Life of Individuality: Modernity, Panopticon, and Dataism	Gal, Tobar, Shokri
Lim, Brian Y National University of Singapore	Designing Theory-Driven User-Centric Explainable AI	Yap, Yang, Singh
Monroy, Raul Tecnológico de Monterrey	Political Bot Detection on Tweets	Dremliuga, Erskine
Singh, Sameer University of California, Irvine	Explaining Decisions of Black-box AI Models	Lim, Shokri, Yang
Shen, Yifan Fudan University	AI Education for Everyone: How to Integrate Future Labor Force into Digital Frontier?	Gal, Dremliuga
Shokri, Reza National University of Singapore	Privacy of Black-box Deep Learning: Analysis and Defense	Monroy, Tobar
Tobar, Felipe Universidad de Chile	How Weak Has Been Weak Artificial Intelligence? The Unseen Societal Consequences of Machine Learning	Lau, Erskine
Yang, Qiang and Toby Tianjian Chen Hong Kong University of Science and Technology	Federated Transfer Learning: Building AI for Everyone with Data Protection	Tobar, Yap, Lim
Yap, Roland	Toward a Certification Framework for Trustworthy AI Systems	Shen, Singh

National University of Singapore		

2 Executive Summary

AI will have a major impact on society. This statement is no longer in question and is consistently reflected across all twelve papers submitted to the "AI for Everyone Project." The current discussion, however, is driven by the notion of whether this impact will be positive or negative. While analyzing possible consequences for different stakeholders, all papers identify and engage with a very practical yet deeply philosophical common theme around the notions of individuality and autonomy, or in perhaps more compelling terms, authority and responsibility for decisions made by machines.

In thinking about this, we must begin with a recognition that much of modern civilization derived by the West has been dependent on an assumption that autonomous individual humans are taking control of, and responsibilities for, consequences of using artifacts. Sustainability of this assumption, however, is under question, particularly in an Asian context that has different philosophical traditions.

Humans will increasingly depend on the judgement of AI, which relies on a vast accumulation of data provided by networked computers. Who is to be held responsible for the accidents caused by autonomously driven cars is a commonly raised question, to which the traditional assumption of autonomous individual human beings does not provide a suitable answer. Protection of individual human rights, including the right to privacy, has become an increasingly complex endeavor as people take advantage of other people's data in many aspects of their lives, including for medical needs. While technical solutions probably exist, we have not yet been able to define what it is that we want to protect.

A venture into the essence of personhood leads us to an interesting proposition, that of recognizing machines as "personalities," at least in a legal sense (as we already legally recognize corporations as pseudopersons). Assuming automated decisions can be attributed to "AI" personalities or institutions, such an arrangement would sound attractive to engineers fearful of being held responsible for all the consequences of AI, including the unintended ones. However, we must emphasize that while acknowledging the increasing popularity of the idea, none of the contributors supported it in the end to avoid evasion of responsibility by others in cases of malicious use of technology.

The notion of machine personality also raises the question of moral agency. Whether or not to allow AI driven weapons to attack people is a real and immediate moral issue we have at hand.

Humans are a heterogeneous species. We must be sensitive to the diversity among cultures, particularly when thinking about the relationship between individuals and communities. The hasty imposition of conventional ethics and standards can cause resentment among the imposed and societal conflicts as new forms of networked communities supported by machine intelligence emerge.

Perhaps less fundamental than the issue of autonomy, but also consistently discussed was the concept of "trust." While there are multiple facets that can be discussed around the concept of trust, two aspects seem especially important. First, humans are wondering if and to what extent technology can be trusted. Second, engineers are wondering what kind of "explainability" they should build into the system so that humans can trust the system.

The concept of trust is characterized by the willingness to rely on another party's actions. When it comes to the relationship between humans and technology, it seems critical for humans to have at least some level of understanding of the technology in order to "trust" and take advantage of the enormous technological opportunity in front of us. At the same time, it seems equally important to build a common understanding of the possible social impacts when building the technology. Thus, technology and new systems should not be developed in isolation but in close communication with other disciplines and perspectives. Put another way, we as a society should think carefully about education, training, and governance for technology development.

In the sea of uncertainties surrounding our future relationship with AI, one thing seems very clear. That is, we must be prepared to work beyond traditional discipline borders to rethink the fundamental elements of our civilization and boldly work together to navigate the technology so that it benefits our future. We must also be prepared to find an adequate response for those who are concerned about malicious uses of the technology.

3 Summaries and Analyses of Individual Papers

This section offers an overview of the submitted working papers, which have been structured based on the four issues identified as societal threats in the context of this project.

3.1 Fear of "black box" machines manipulating human society

Unlike traditional information systems that essentially operated along predetermined algorithms, much of AI's behavior depends on the data that it learns from and is not transparent to the users. Papers have been written on this subject from both a technical standpoint to identify and solve the problem as well as from a behavioral perspective, analyzing how humans may in fact be manipulated. Delegation of monitoring to others, such as experts, by means of certification may also provide a solution.

Lim (2018) discusses the importance of "explainable artificial intelligence," which allows end users to understand the models and algorithms at work in systems, leading to better control. To this end, the author developed a framework of explanation layers that addresses the question of why people make certain inquiries and what the main goals for explanations should be. Methodologies on how to provide explanations were developed based on the stated goals. People's cognitive limitations are analyzed to propose a "user-centric explainable AI" that reduces decision errors by users (in this paper, medical professionals).

Singh (2018) discusses the prospect of increasing human trust in machines by providing a model-agnostic and intuitive way to explain any machine learning algorithm. Successful development of such a method would allow humans to understand the behavior of AI, reducing their perception as "black boxes." The paper focuses on the explainability of classification models, on providing interpretable descriptions of how input affects predictions. It also sets goals in providing "local" explanations of why a specific decision was made for a specific instance. Three forms of explanations, namely linear models, anchors (sufficient conditions), and counter examples are introduced. Illustrative applications of these three forms are provided.

Gal (2018), taking on the case of Replika, offers empirical evidence on how human users of chatbots can build emotional attachments to the pseudopersonalities AI generate. By using a variation of the Uses and Gratifications Theory to guide a structured content analysis, the author analyzed 447 user reviews of the application. The analysis indicates that using Replika is a largely gratifying experience, especially when looking for artificial companionship. It also finds, however, that engagement in this artificial companionship can negatively affect the users, both emotionally and socially. Implications of this "artificial socialization" are discussed alongside a call for continued study of the topic.

Yap (2018) offers a framework for the development of a certification system for trustworthy AI systems. The author identifies such issues around AI as: (1) how to show that the results are correct or accurate, (2) how to explain and make interpretations on the results, and (3) how to show that the results are fair. The author then explores how people can put their trust in technology under such limitations, concluding that to be trusted, systems have to (a) be developed by trustworthy developers, (b) be fair, and (c) generate understandable results. They also need to be secure, even in an adversarial environment, and come with assurances. "Certification elements" are proposed to develop a certification system to increase trustworthiness. Such proposed certification elements should be designed to include technical expertise as well as neutrality in their evaluation. Another option presented is to allow self-certification, relying on disclosure by system providers.

3.2 Recognition that AI may be put to unethical uses and that some restraining mechanisms are necessary

Malicious, unethical use of AI can be catastrophic to human existence. This part of the project

studies moral, technical, and legal solutions to this problem. Analysis is focused on specific contexts, such as war and political manipulation, using online robots. More fundamental questions around responsibility are also explored from the perspective of criminal law.

Erskine (2018) focuses particularly on conflict situations, namely war, to argue the importance of the notion of "moral agents" for "moral restraints." Moral agents are defined as "actors that possess capacities for understanding and reflecting upon moral requirements, and for acting in such a way as to conform to them." Comparing flesh-and-blood humans, corporations (institutions), and AI, only humans (and institutions, to the extent individual humans can be held accountable for them) have the quality of being "moral patients" vulnerable to suffering from a breach of morals by others. Given that robots lack this quality, it seems appropriate for humans to be the ultimate moral agent. The author points out two kinds of risks involved in allowing non-flesh-and-blood agents to be moral agents as follows: (1) abdicating responsibility to non-moral agents and (2) eliding the responsibilities of distinct moral agents.

Monroy et al. (2018) reviews the literature on - and offers views on - bot detection mechanism. Botnets can be used to provoke trending topics and have been proven to be effective in either favoring political figures, misrepresenting said figures' opponents, or influencing voting behavior. Therefore, detection of botnets is of high importance. Bot detection on Twitter is based on the premise that genuine client accounts show different behavior to bot or semi-automated accounts. By analyzing combinations of tweet content, sentiment, tweet account, account usage, and social network features, detecting the existence of botnets becomes possible. The authors take a contrast pattern-based approach to detecting botnets. They are also testing a "generative adversarial network" approach. Results are forthcoming.

Dremliuga (2018) questions the future viability of our present day criminal law framework. The spread of social systems where humans and AI coexist increases the ambiguity surrounding intentionality and culpability in harmful events. Hence, criminal law systems that focus on penalizing responsible individuals may become neither fair nor effective at preventing harm. The author moves on to discuss the practicality of giving legal personhood to machines in a manner analogous to corporate personhood. While recognizing increasing support for the idea, the author questions its effectiveness, as it may fail to penalize and deter the malicious use of technology. In conclusion, the author proposes a revision of the concept of culpability, suggests increased control of the possession of powerful AI systems, and opposes the adoption of the machine personhood concept.

3.3 Risk of inference attack on privacy, i.e., breach of privacy through AI analyzing the results of predictions to determine attributes of subjects or databases

While privacy has been an issue from the early days of Internet use, the issue is taking on a

new and more serious character given AI's great power of inference. That is, by identifying and learning patterns in data, machines are capable of "mining" sensitive private information from data and/or even from results of the predictions generated by other systems. A philosophical revisiting of the notion of the individual is also necessary for a more fundamental understanding of the meaning of the emerging term "panopticon."

Shokri (2018) points out the dangers of "inference attacks," i.e., breaches of privacy resulting from AI analyzing the results of predictions to determine the attributes of subjects or databases. The author then seeks algorithms to minimize the threat of inference attacks while maximizing predictive capabilities of AI at work. This is done by introducing a "regularizing" function designed to prevent the predictive model from overfitting. Experiments were conducted using adversarial inference attack models to test the effectiveness of the concept.

Yang and Chen (2018) explore ways to integrate data that is scattered across various organizations while protecting the privacy of the individuals. The authors propose three types of implementations, namely (1) transfer learning, (2) federated learning, and (3) federated transfer learning, where "transfer" involves the transfer of knowledge (not data) and "federated" refers to integrating data on the same subject in different parts of the schemas. The authors point out data waste in the conventional federated approach and recommend a federated transfer learning approach.

Lau (2018) explores the notion of individuality in the context of AI, seemingly providing a contemporary realization of "panopticon," or a system of governance by ubiquitous social surveillance. The author discusses how the West has been developing the concept of the individual's freedom as a human right, but quickly adds that a recognition of resulting conflict among the interests of the individuals has brought about a notion of "disciplining of individuality." The idea of panopticon emerged in this context with the goal of providing maximum surveillance with the least effort. Both governments and large commercial online services are in a position to play the role of the "inspector" and are becoming a contemporary threat to the notion of individuality. The difference between the current and the nineteenth century context is that nowadays, people are often the beneficiaries of such surveillance, leading them to be willing participants. The invisibility of surveillance is also a characteristic of contemporary times. Thus, the all-encompassing data system neither suppresses nor controls any individual, but it dissolves individuals into a new collectivism that is made up not of individuals but of something more fundamental than humans. We may be heading into an age of posthumanism or transhumanism.

3.4 Fear that AI may widen the gap between the rich and the poor

Changes in the world of work and especially loss of jobs has been one of the primary concerns in the context of discussing AI. While fear of job loss is nothing new in the interaction between technology and society, AI's flexibility in performing highly contextual work adds a whole new dimension to the problem.

Tobar (2018), with “replacement of humans” as an underlying theme, identifies two lines of developmental philosophies for AI. One pushes AI researchers to concentrate on examining whether machines can have a mind (Babbage approach). The other focuses on the replacement of humans by AI while recognizing the impossibility of artificially creating a mind (Cartesian approach). The author states that this categorization is in line with Searle’s distinction between strong and weak AI. Weak AI researchers have shown that “weak AI has not been as weak as was originally thought.” While a rule-based approach has limits, machine learning can result in machines simply replacing human labor.

Shen (2018) focuses on people born in or before the 1990s in China, as they are considered less trained in AI-related fields and thus highly susceptible to technological advances in the job market. The author provides empirical evidence of a trend toward increasing wages and a labor force shift in China between 1978 and 2016, coinciding with the introduction of computers and automated machinery. Given this evidence, the author proposes a list of college majors that need to incorporate computer science education, and provides an assessment outline for the design of AI courses offered by both colleges and corporations.

Policy Statement

- AI (artificial intelligence) has the potential to benefit humankind enormously if it is governed adequately. However, it also has the potential to harm humanity catastrophically if it is mismanaged.
- This message, coming from the Pacific Rim Universities, particularly emphasizes the need for sensitivity to the diversity in culture, religion, and political systems when developing governance philosophies and structures.
- Governments, academia, businesses, and non-profit organizations must work together across cultural and political boundaries to establish trust in technology, both by adequately managing the technology and by enabling people to use the technology to beneficial ends.
- The Association of Pacific Rim Universities, with its members coming from a diverse range of political and cultural backgrounds but united behind its academic rigor, offers a unique platform for open discussions.
- In this project, aimed at building a community of researchers on the beneficial use of AI, we have been successful in agreeing on a common goal, that is, *access to the benefits of AI, awareness about the nature of the technology, governance of the technology, and its development process with a focus on responsible development should be transparent, open, understood by, and accessible to all people regardless of their geographic, generational, economic, cultural, and/or other social background.*
- We have also identified the following major issues to be addressed:
    - *Fear of "black box" machines manipulating human society*
    - *Recognition that AI may be put to unethical uses and that some restraining mechanism are necessary*
    - *Risk of inference attack on privacy, i.e., breach of privacy through AI analyzing the results of predictions to determine attributes of subjects or databases*
    - *Fear that AI may widen the gap between the rich and the poor*
- AI is likely to change the foundational constructs of human society, such as autonomy, ownership, and markets. While using conventional norms to manage immediate issues, we must be prepared to think out of the box to offer alternatives regarding the future of humanity.
- While the research is still very much preliminary, we are actively pursuing opportunities to interact with policymakers, businesses, and leaders in society to exchange ideas based on substantial scientific evidence and constructs that reflect history and cutting-edge technologies.