Global AI Dialogue Series
Takeaways from the Asia Pacific-US AI Workshop: Measuring Impact, Building Inclusive AI, and Bolstering Trust in the Digital Ecosystem

May 9, 2018 Hong Kong SAR, China

Berkman Klein Center for Internet & Society at Harvard University and the Digital Asia Hub, in collaboration with United Nations University Institute on Computing and Society & The China Institute for Science and Technology Policy at Tsinghua University

Workshop Rapporteur: Jenna Sherman[1]

I. INTRODUCTION

The Asia-Pacific region and the U.S. are pioneers and early adopters of numerous innovations in the ongoing AI and automated systems technological revolution. However, at the same time that the respective regions are targeting high growth and investing heavily in AI, new governance challenges ranging from privacy and security to unbiased decision-making are also rising. The promise of AI-based technologies is enormous -- private firms and public institutions will benefit massively from AI-enabled efficiency gains and other unprecedented improvements across sectors -- but barriers to these gains and potential externalities are equally significant.

In order to establish a cross-cultural dialogue and learning network on specific AI issues and potential methods for addressing them within and across the US and APAC region, the Asia Pacific-US AI Workshop was convened by the Berkman Klein Center for Internet & Society at Harvard University with the Digital Asia Hub, in collaboration with the UN University Institute on Computing and Society and the China Institute for Science and Technology Policy at Tsinghua University. The objective of the meeting, which convened 35 subject matter experts from across academia, industry, government, and civil society, hailing primarily from several APAC countries as well as the US, was to provide a platform for stakeholders with policy, business, or technology responsibilities to develop and share insights on AI from an ethics and governance perspective. The meeting encompassed three thematic tracks, 1) AI Indices: APAC Data, the China AI Index, and the AI Index, 2) Measuring AI's Social Impact - State of Play and Looking Forward, and 3) (Re-)Establishing Trust in the Digital Ecosystem, all of which were explored through a range of full-group discussions, breakout working sessions, and report-backs that included input statements and case study presentations from leading contributors.

This write-up seeks to share observations from the workshop, highlight overarching themes that emerged, and extract insights on next steps for sustaining the cross-cultural dialogue and building out from it. The distilled outputs are centered around five key takeaways that emerged throughout the workshop sessions:

---

[1] Project Coordinator, Berkman Klein Center for Internet and Society, jsherman@cyber.harvard.edu.
Special thanks to Nathan Kaiser, Fellow at the Berkman Klein Center and co-organizer of the workshop, Amar Ashar, Assistant Director of Research at the Berkman Klein Center, and Ryan Budish, Assistant Director of Research at the Berkman Klein Center, for their inputs on this write-up.

1. In order to paint a robust picture of AI's societal influence, the lens must shift from market to impact metrics
2. There is an emerging awareness that despite some applicable lessons from previous technologies, AI is fundamentally different
3. Responsibility for mitigating AI risks lies in a triangulation between users, industry, and government
4. Discussion pertaining to trust and ethics in AI should not be divorced from similar discussions in other technical areas
5. Despite the lack of consensus on a best way forward, all players and stakeholders should continue to actively iterate on interventions that seek to bolster trust and ethics in the digital ecosystem

These key takeaways were distilled with the intention of serving as a heuristic tool that allows us to map the most salient outcomes of the workshop, the context from which they stemmed in the meeting, and how they might be implemented. However, it is worth noting that there is significant overlap across and within these categories, and that this write-up is intended to represent a snapshot of the APAC-US AI Workshop, situated within a broader conversation on AI and its societal impact.

II. KEY SESSION TAKEAWAYS

1. In order to paint a robust picture of AI's societal influence, the lens must shift from market metrics to impact metrics

There are a number of initiatives underway tailored towards collecting, measuring, and distilling data on AI from a comparative, international perspective. During the workshop, participants working on a variety of these initiatives showcased their efforts, including the China AI Index, conducted by the China Institute for Science and Technology Policy at Tsinghua University, and the 2017 AI Index Report, and explored ideas and opportunities for future collaboration.

Representatives' presentations and the subsequent group dialogue sparked agreement among participants that in order to begin to understand AI's societal footprint, particularly from an inclusion angle, there is a need for metrics to shift from ones pertaining to the AI market -- which employ a primarily quantitative approach to measure the rise in the development and adoption of AI -- to impact metrics -- which use a mixed-methods approach to analyze how AI impacts society. Reaching this accord required first an understanding of the current measurement efforts and indices at play, followed by group dialogue.

Current Indices, Metrics, and Approach
Representatives from the AI Index, and the efforts at Tsinghua University to collate a China Index, titled the "China AI Development Report 2018," offered an overview of the current methods and findings from their measurement efforts thus far.

The 2017 AI Index Report, a project within the Stanford 100 Year Study on AI (AI 100), is an initiative to measure AI's impact on society, the growth of AI, and to better understand AI's future trajectory. The culmination of the initiative's work, thus far, is a comprehensive 2017 report aimed at tracking and analyzing data on AI as a resource for various stakeholders, such as policymakers, NGOs, and researchers. Currently, the Index is seeking to expand its inquiry into more global AI contexts, and to combine or coordinate efforts of analyzing AI growth and impact. In an effort to do so, two next steps of the endeavor include identifying application-specific growth in the AI space - such as healthcare in AI - and improving the indices selected for exploration.

Topics covered within the China AI Development Report 2018, as presented by a Tsinghua representative, include the intersection between AI and education and AI and health, explored primarily via questionnaires and metrics analysis, such as analyzing trends of keywords related to AI in course catalogues and examining which universities were offering relevant courses at higher and lower rates. A primary objective of the report is to present snapshots of the AI ecosystem in China from different use cases. The education use case shed light on how AI is perceived and adapted by universities, the opinions of those who have taken AI-related courses, and the demographic variables of those taking the courses such as age, gender, and education level.

Samples of key findings from the research thus far include that the majority of these courses are offered by the school or department of computer science in universities -- as opposed to math or social science departments, for instance -- and that Coursera is the most popular platform utilized for online AI-related courses online. Additionally, the majority of enrollments are among men, in line with skewed gender demographics typical in the STEM field. Findings such as these help paint a more robust picture of the current status of AI education and application in China, which in turn can help shape AI design, implementation, evaluation, and education efforts to be more inclusive and beneficial on aggregate.

Moving from Understanding the Ecosystem to Impact Measurement

Project representatives and attendees alike discussed moving beyond metrics-based analyses of understanding the AI ecosystem to impact-centered measurement. Adopting this approach would entail conducting an analysis of the ways and extent that AI impacts society at a global level, particularly across race, geographic location, and socioeconomic status, rather than chiefly measuring the rates at which AI is developed and studied. Though this endeavor would necessitate a more nuanced approach, as there is high variance in the ways in which impact can be assessed, attendees discussed how having data on both AI production and AI impact is integral for considering targeted interventions, solutions, and collaborations to leverage AI for the social good and desilo who the producers and beneficiaries of AI-based technologies are.

Turning towards this type of analysis might also aid predictions of how AI will continue to grow and its future impacts. One posited example involves examining how the impacts of increasing open source access in South Korea might alter the AI ecosystem within the country moving forward, as open source is currently lagging in Korea. Another proposed example included measuring how much training data is available to what groups of individuals, and inferring how inclusive AI will be, or already is, for those who it's supposed to serve considering that datasets are often siloed.

In embarking on this impact analysis, there is also room to incorporate an increased number and range of metrics, research methods, and questions. Specifically, using social science and policy factors as indicators in an effort to complement more quantitative measures of AI primarily analyzed now, such as surveys, patents, and startups.

Opportunities for a Global Indices and Limitations

Within the broader suggestion of increased impact analysis, a number of tangible opportunities emerged. In regards to the AI Index, the project leads discussed the merits of executing a finer-grained analysis of how AI is impacting the economy and promoting regional growth and innovation, which may involve modeling methods that help diversify AI and equitably expand its impact. The project also discussed the importance of adopting a more international lens for analysis and examining how we model the supply and demands for AI data trends within academia to yield global AI metrics.

In the same vein of expanding cross-collaboration on AI metrics analysis globally, one potential nexus that emerged is the possibility of Korea University Law School conducting a more tailored measurement of societal impact in Korea and the APAC region more broadly, as opposed to metrics on development and market impact, in line with the broader suggestion of transition outlined above.

Concrete models and tools for sharing data and research methods also play a critical role in yielding the global and multi stakeholder cross-pollination discussed, so as to encourage the desiloing of data and encourage a more robust understanding of the global AI ecosystem. Participants discussed low-hanging fruit that could fill this role and serve as both a tool and connector. One idea involved the creation of a neutral, cross-sector database with an academic focus meant for sharing and circulation as a "living" repository for data storage and analysis. Such a database could be significant in pinpointing the correct data required to analyze varying metrics, such as AI development and growth versus impact, and would potentially help to foster a greater understanding of AI's impact on an international scale. Such an exercise could be useful in mapping governance and policy strategies, as well as understanding the variance of local impact, and may organically lead to collaborations and outcomes such as analysis reports.

2. **There is an emerging awareness that despite some applicable lessons from previous technologies, AI is fundamentally different**

   Participants acknowledged the importance of understanding the ways in which AI differs from preceding technological developments, such as the automobile and the internet. While it's challenging to come to a consensus on this question, participants felt it was imperative to base conversations on AI impact in open dialogue about the uniqueness of AI both inherently and when contextualized, prior to exploring its governance and regulation. This grounding was expressed as necessary, as AI's relation to other technologies influences the lessons we can learn from and apply to AI from those prior technological developments and adoption.

   What Makes AI "Different?" -- Moving from Today's AI to the Future
   Stemming this agreement among participants, there was an undercurrent of questions throughout the workshop on the very nature of AI and how we ought to be conceptualizing it. For instance, one participant inquired to what extent AI a "reflection of what is," rather than what ought to be? As in, how much of AI is simply a parallel and amplification of our own beliefs and dynamics, and how much is it generating new patterns and new information? A more pointed question that stemmed from this broader one asked that if data is biased, is the technology to blame or are we to blame? One attendee postulated that we are likely to find AI to be similar in some respects to technologies that have come before, and that our measurement of AI can, in fact, be deeply informed by how we measured the impact of these prior technologies, giving us significant grounding to stand on.

   A lesson from all pivotal innovations discussed is that AI will always reflect the values of its creators, from who sits on a corporate board to who the actual designers are, and that without intentionally fostering an inclusive environment of AI developers and stakeholders, we risk building AI that mirrors narrow and exclusive perspectives. One participant noted that in a study conducted on gender representation in work on AI, women compose just 8% of team members among AI companies and startups, and 22% of the top academic institutions working in this field -- a clear reflection of inequality in this sphere, which in turn is reflected in AI and its outcomes.

   Participants also identified ways in which AI differs from preceding technologies, and why, though there are overlaps, these require new forms of auditing and overall regulation that diverge from previous approaches. The case of the automobile sheds light on the discrepancies: when you're driving a car, you're not impacting how the car operates internally -- it operates uniformly regardless of who's driving it, and does not alter based on the operator's behavior; this is not the case with AI, as when your data is involved, you are inherently part of how the technology works, and are yourself involved in its development. Therefore, regulation must account for this learning component involvement of individuals and entities in the fabric of the technology itself. An additional

identified fundamental difference between AI and other technologies, stemming from this learning mechanism, is that AI may “develop” opinions that in turn shape our opinions, rather than simply operating as a stagnant tool to serve us. Therefore, greater thought and intentionality is required to put into how we’re training and “parenting” our AIs, attendees discussed, as Strong AI, or Artificial General Intelligence, is going to eventually emerge from the current system of weak AI, which should influence our approach to development and regulation of AI differently than previous examples.

Building a Taxonomy for Measuring Impact

In considering the discrepancies between AI and other technologies, participants touched on the value of identifying a new, or adjusted, taxonomy for measuring impact that moves beyond the typically-employed approach of measuring pre-identified metrics, recording the numbers, and quantifying impact.

One proposed approach for this taxonomy, warranting further exploration, follows a layered model based on the complexity of the algorithm, as well as its risk. The framing starts with explainability as the most sought after version of knowability -- which involves furthering the understanding of the algorithms themselves and having a firm grasp on its technical decision-making processes. Following explainability is auditing, which can probe the algorithm to test its outcomes as a means of sound understanding, though the algorithm itself may be a “black box” to us on the technical end. The third tier is knowability, which suggests that so long as the AI-based tools are yielding fair and inclusive outcomes, it’s usable, but should not be if it doesn’t meet each of those markers. And finally, the proposed bottom layer is redline knowability, which proposes that if the AI is a critical decision-making system, its processes must be subjectable to moratoriums, given the higher room for error and negative impact.

3. **Responsibility for mitigating AI risks lies in a triangulation between users, industry, and government**

Within any impact taxonomy, however, there must be means for mitigating the broader risks of AI proactively, in order to leverage the beneficial impacts and reduce risks. Participants agreed that to do so requires a triangulated approach that serves to simultaneously encourage powerful stakeholders in government to promote effective and beneficial AI regulation, incentivize industry to develop more inclusive and less risky AI, and empower individuals to take agency over their own AIs and foster a healthier AI-powered ecosystem.

A few broad-level suggestions include considering a type of impact measurement that allows us to measure direct impact versus indirect impact. This type of impact measurement would likely involve some of the aforementioned measures, but might also encompass more grassroots data collection and longitudinal studies that observe the

trajectory of different facets of society -- such as healthcare, education, and labor markets -- in conjunction with the growth of AI.

Additionally, participants discussed the need to consider who the key stakeholders are in industry and public sectors, and demand algorithmic transparency and accountability, as individual input is limited. Doing so might entail requesting a standardized auditing process, building an open database of algorithms used by various platforms, and devising systems for users to voice questions, concerns, and complaints about the tools. One concrete approach suggested by participants for governments and industry is to conduct comprehensive testing of AI tools to identify the level of error rates across different indicators, examine why those error rates are occurring, and adjust for the issue accordingly.

Individuals do, however, still have a role to play. Though we are not yet at the point of Strong AI, or Artificial General Intelligence (AGI), the AIs of today are the foundation of the AIs of tomorrow; therefore, users need to put more thought and intentionality into how they train and "parent" AIs. Doing so not only has implications for our own digital systems and the ways in which our own AI-based tools learn, but also for the algorithms that are built off of it, and the environments in which our AI systems are interconnected -- such as search engines and social media platforms that influence our lives. When considering this type of AI "parenting," particularly when it comes to social media and search engines, one attendee proposed that individual users should be thinking about improving AI in three primary areas, in conjunction with proactive efforts by government and industry:

1. Diversity of opinions: as AI inherently creates echo-chambers by learning by only what content we want to interact with.
2. Respectful open-mindedness: as AI often rewards trolls and hateful comments as they get more views.
3. Sound judgment: as there is currently no system of flagging content or AI-based decisions that seem questionable or fake.

4. Discussion pertaining to trust and ethics in AI should not be divorced from similar discussions in other technical areas

Throughout the engaged discussion at the workshop, it became evident that questions at the intersection of ethics, values, and AI can be applied to other facets of our digital ecosystem, including content regulation, harmful speech online, data collection, and online surveillance -- and that these underlying ethical questions we're exploring related to AI should, in fact, be explored in tandem in order to get at root answers that apply across technologies. Though each of these components requires a unique analysis and intervention, there are broader, fundamental questions threaded throughout each -- and

exploring those questions in the context of one issue can help translate to answers for another.

Participants additionally discussed how education and bridge-building across disciplines and problem areas is key to improvements and progress in the field of technology and society. For instance, one attendee suggested that lawyers knowing only law is not the future, rather, the future is interdisciplinarity given the interconnectedness of the globalized and technologically-infused world. The inability to do so, they argued, would stunt our progress for more just and equitable technologies due to a lack of practitioners able to tackle challenges across sectors, while the technology continues to evolve.

Participants also tackled questions surrounding fairness, transparency, and accountability in the digital ecosystem, and advanced the idea that it is often when those aspects are absent that a lack of trust is cultivated. Exploring where the gaps in these three pillars exist, where they stem from, and how we can begin to adjust for those gaps are first steps for constructing increased trust and, in turn, increased benefits for users.

There needs to be some understanding of, for example whether a credit system can use one's data to influence someone and how they're treated, such as in differential insurance pricing and non-transparent ways in which users are pooled and profiled. In this same vein, we need to be exploring whether the data being collected are appropriate data, and how its collection impacts the user, such as in the advertisements one is served or the price of an airplane ticket you're offered. Within each of these examples it's evident that in discussing ethical technologies, we are in reality discussing human ethics, and it's therefore our obligation to determine standards and best practices from how we can opt out of data tracking, to how content is regulated, to how autonomous vehicles will respond in crises.

5. **Despite the lack of consensus on a best way forward, all players and stakeholders should continue to actively iterate on interventions that seek to bolster trust and ethics in the digital ecosystem**

One key question that surfaced through this dialogue was if trust is a cycle that might evolve, and, for instance, will there be more trust and autonomy in the system going forward as it develops in parallel with our understanding of its functionality? Autopilot, for instance, is now set to take over for a pilot in certain emergency circumstances, as it's proven to make the more effective decision. This improved decision-making, as well as the increased trust required to implement such an autonomous system increased over time, perhaps sheds light on the ways in which trust in AI and the digital ecosystem may evolve. Regardless, measures throughout the process of reaching that level of trust are not only helpful but imperative, as without testing different approaches, even if ultimately ineffective, it is impossible to ever reach that point.

Participants shared a number of concrete ways in which they are involved with efforts to augment trust in the digital ecosystem through various modes of intervention, such as technological improvements, new design techniques, policy implementation, and altering social norms.

On the technological level, a representative from the US industry spoke to how their company recently made tweaks to their autocomplete algorithms to remove any results suggestions that signal authoritativeness on the search engine. In addition, they're working on creating signals for users to help fact check online information presented in search results. In regards to hate and harmful speech online, the same company has also begun using algorithms to flag harmful speech on social media, and is pushing forward greater digital literacy efforts to discourage the popularity of harmful or falsified video content, particularly among youth who comprise a significant portion of the platform users.

Within the sphere of policy and governance, a telecommunications company in China is looking for advisories on how to move forward with improving content quality on platforms, bearing in mind that different countries have their own unique problems. By implementing advisories, solutions for tackling effective and fair content regulation have the potential to be applied more uniformly. The same American industry representative also spoke to how the company has implemented new policies to dis-allow harmful autocomplete suggestions in the engine's search function. With this approach, the content in some cases, when not in direct violation of the platform norms and policies but nonetheless potentially harmful, can remain online but with decreased accessibility in that the algorithms will not suggest that content or promote its popularity. Finally, another participant spoke to how the Personal Data Protection Commission (PDPC) in Singapore sets a precedent in this field, and promotes the legal protection and ownership of personal data across domains such as health, finance, and education.

Finally, with respect to norm-shaping and education, one contributor spoke about a holistic law and technology education program being implemented in Singapore. The program does not provide certification to practice law, but provides an interdisciplinary education at the intersection of technology and the law, which, this individual articulated, is a necessary and valuable program as AI and other technologies continue to develop. Through this program and other measures, the Singapore government is trying to establish greater trust in the digital ecosystem as they prioritize safety.

In a separate endeavor, a Japanese representative discussed the conversation surrounding AI and ethics shifted regionally when an AI journal cover displayed a female cyborg-servant as the pinnacle of the AI future. Not only did the national dialogue begin to correspondingly shift, but in response, the Japan Society for AI implemented their first set of ethical guidelines for AI. Since then, the Japanese Ministry of Internal Affairs and Communications released R&D guidelines for AI, and now in Japan a number of groups

are forming the board to review these guidelines and their implications. In a more policy-oriented approach, The University of Tokyo RIKEN Center for Advanced Intelligence Project conducted a technology assessment report geared towards policymakers. This use case demonstrates how a shift in norms can influence policy, and vice versa, and how one without the other runs the risk of resulting in a continuation of the residual lack of trust.

II. CONCLUSIONS

Throughout the participatory workshop, as dialogue shifted from measuring the impact of AI to practical methods of mitigating AI risks and reestablishing trust in the entirety of the digital ecosystem ecosystem, participants agreed on the need for greater exploration of the broader questions surrounding ethics and technological impact, with simultaneous proactive work on practical methods of mitigating risks and maximizing opportunities. Within this dual-approach, attendees emphasized potential action items, including conceptualizing projects related to AI indices, with an emphasis on inclusion metrics, implementing taxonomies for fairness and impact measurement, identifying quantifiable paradigms, and following up on tangible, actionable next steps across sectors and regions to build upon the momentum catalyzed at the meeting and further partnerships.

In the same vein, participants agreed that it was imperative to continue fostering a cross-cultural dialogue between the Asia Pacific Region and the U.S. on the impact of AI across regional contexts, particularly as informational and methodological asymmetries remain. By continuing these conversations, it becomes increasingly possible to create a more inclusive roadmap for meaningfully and effectively measure AI's social impact, informing an AI global governance framework for international policy-makers to promote beneficial AI, and increasing trust in the digital ecosystem by implementing measures that yield a healthier, more robust technological landscape and future.