April, 2021

AI Principles to Practice

An Introduction to International Activities

Danit Gal The University of Cambridge

Tina M. Park Partnership on AI

Yolanda Lannquist The Future Society

Adriana Bora The Future Society

Niki Iliadis The Future Society

Mia Shah-Dand Women in AI Ethics™

Olga Afanasjeva GoodAI

Arisa Ema The University of Tokyo

Table of Contents

03 Editor's Introduction to "From AI Principles to Practice: Introducing International Activities"

Arisa Ema
Institute for Future Initiatives, The University of Tokyo / RIKEN

06 Japan's Role in Multilateralism on AI Governance

Danit Gal
Leverhulme Centre for the Future of Intelligence, the University of Cambridge

13 Paving an Intentional Path Towards Inclusion in AI Development

Tina M. Park
Partnership on AI

20 AI for Good and its Global Governance

Yolanda Lannquist
The Future Society

Adriana Bora
The Future Society

Niki Iliadis
The Future Society

28 The Diversity and Ethics Crisis in AI

Mia Shah-Dand
Women in AI Ethics™

36 The greatest scientific challenge

Olga Afanasjeva
GoodAI

39 Commentary: From AI Principles to Practice: Lessons for Japan to Learn from International Activities

Arisa Ema
Institute for Future Initiatives, The University of Tokyo / RIKEN

Special Issue "From AI Principles to Practice: Introducing International Activities"

Arisa Ema Institute for Future Initiatives, The University of Tokyo / RIKEN

From Principles to Practice

The title of this special issue, "From Principles to Practice," is the title of the first chapter of "Ethically Aligned Design, 1st Edition," published by the Institute of Electrical and Electronics Engineers (IEEE) Global Initiative on the Ethics of A/IS in March 2019 [1] . Currently, numerous principles and guidelines on AI ethics have been published. However, it is difficult for people outside the AI community to understand these guidelines, the organizational efforts being made to implement these guidelines, and the major stakeholders.

International and multistakeholder discussions are considered important. However, the fact that there are various fixed stakeholders is criticized. I have attended several international multistakeholder meetings and observed that the same group of people are present in a few cases. In addition, not so many Japanese people are now participating in international discussions compared of that of the West countries.

This special issue requested those who are leading/working on the formation of AI ethics and governance communities and networks worldwide to introduce their activities, current efforts and challenges, and their expectations for Japan and the Japanese research community.

Structure of this special issue and introduction of authors and organizations

In the commentary at the end of the issue, I have introduced the lessons from each article that are relevant for Japan. Here, I briefly introduce the authors of this special issue and their activities.

The first author, Ms. Gal, is currently with the University of Cambridge. She is a part of various international and interdisciplinary networks. She serves as the vice chair of the P7009 IEEE standard on the Fail-Safe Design of Autonomous and Semi-Autonomous Systems. In addition, she is a member of the executive committee of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, which published "Ethically Aligned Design." She is an executive committee member of the AI4SDGs Cooperation Network at the Beijing Academy of AI [2] . She is a valuable person who is well informed about the West and East contexts of AI. Until the end of December 2020, she was a Technology Advisor and AI Lead at the Office of the Under-Secretary General and Special Adviser on Preparations for the 75th UN Anniversary & the SG's Roadmap for Digital Cooperation. She was a special guest at the 2017 Annual Conference of the Japanese Society for Artificial Intelligence (JSAI), "Open Discussion: The Japanese Society for Artificial Intelligence [3]."

The second author, Dr. Park, is currently leading the Methods for Inclusion project at the Partnership on AI (PAI). The project has been recently launched, and it is actively engaged in research and activities on AI. The PAI was initially founded by IT giants such as Google, Amazon, Facebook, DeepMind, Microsoft, and IBM. However, the majority of current members are nonprofit organizations, and the official website [4] reports that there more than 100 partners from 13 countries. The members from Japan are Sony, the Next Generation Artificial Intelligence Research Center of the University of Tokyo, and Softbank. PAI conducts research on specific issues such as AI fairness, impact of AI on future work, and safety of AI.

The third author, Ms. Lanquist, and her colleagues belong to The Future Society, which is another nonprofit organization that has been gaining prominence in recent years. This organization was founded in 2014 as a think-and-do tank at the

Harvard Kennedy School. In recent years, it has been working with international organizations and governments on AI policy research, educational projects, and advisory projects. It has led the research conducted by the "Responsible AI" and "AI's Pandemic Response" working groups of the Global Partnership on AI (GPAI), which is an international and multistakeholder initiative by France and Canada to discuss issues related to AI. Prof. Osamu Sudo (Chuo University), who chaired the committee that formulated "Social Principles of Human-Centric AI," and Prof. Toshiya Jitsuzumi (Chuo University) are participating in this GPAI working group from Japan.

The fourth author, Ms. Shah-Dand, is the CEO of Lighthouse3, which is a research and advisory firm in the field of AI utilization, and a founder Women in AI EthicsTM. She proposed the idea of supporting women working in the AI ethics and governance area. In 2018, she single-handedly selected and published the first "100 Brilliant Women in AI EthicsTM" list, which was a resounding success. Currently, she is supporting events for women and minorities who are active in the AI ethics and governance area and launching a support program for students and researchers who require help in dealing with COVID-19.

The fifth author, Ms. Afanasjeva, is the COO of GoodAI, which was founded in 2014 by Marek Rosa with an investment of $10 million with the goal of developing artificial general intelligence (AGI) to help humanity and understand the universe. AGI would provide tremendous benefits if it is realized. However, the development, operation, and usage of such a system must be carefully discussed. With this objective in mind, GoodAI is organizing the "General AI Challenge" to solicit wisdom and knowledge on how to prevent the occurrence of an AI race. Ms. Afanasjeva and Mr. Rosa were speakers at the AI and Society Symposium held in Tokyo in 2017 [5], and it was exciting to hear about the results of the General AI Challenge, which was still in the conceptual stage at the time.

Multistakeholder discussions are essential for addressing the various challenges of AI. This is demonstrated by the fact that the affiliations of the abovementioned authors are diverse, including academia, international organizations, nonprofit organizations, and corporations. This suggests that organizational fluidity and a wide range of activities are important for resolving problems such as those discussed in this special issue.

However, it should be noted that owing to time constraints, the authors were mainly from Europe and the U.S. and limited to the editor's acquaintances. The discussion on AI is being led by excellent people in the Middle East, Eastern Europe, Africa, East Asia, Southeast Asia, and Oceania, and the diversity of discussion subjects is increasing.

In terms of diversity, you may have noticed that all the authors of this special issue are women. In the planning of this special issue, most of the candidates were women. Thus, as an editor, I decided, based on my own judgment and preference, that the issue should be composed entirely of women authors. Numerous women are active in the field of AI ethics and governance worldwide, including those on the "100 Brilliant Women in AI EthicsTM" list. As a result, from the viewpoint of diversity, the composition of this special issue was biased. However, I hope that in the future, we will be able to ensure the diversity of authors in the journal without bias.

Conclusion

I would like to express my gratitude to those who contributed to the compilation of this special issue. First, I would like to express my special thanks to the authors for agreeing to write this article in a short period of time (approximately a month), over the year end and New Year holidays. I believe that their willingness to do so was due to their expectations for Japan. As mentioned at the beginning of this article, among the various challenges posed by AI, the debate on diversity and inclusiveness is a major global issue. Various organizations have been criticized on the lack of diversity among participants in "conferences that emphasize diversity and inclusiveness," including Japan's "Social Principles of Human-Centric AI," in which only 13.8 % members are women. Particularly in international discussions, where most of the participants are from Western countries, there are surprisingly high expectations of participation from Japan, an island nation in Asia and the world's third largest economy and technology powerhouse.

Owing to time constraints, I translated the entire Japanese

text to English in a week (a few hours every night and on weekends). The original text was put through several machine translation services, and the technical terms were translated based on my knowledge and experience, articles and papers, and voice recognition software in a few cases. In a sense, it was a collaborative effort between humans and AI. I would like to express my sincere gratitude to Mr. Kiyota, the editor-in-chief, for his time and effort in not only formatting the manuscript but also checking and correcting expressions in Japanese and references.

We hope that this special issue will provide a springboard for more people involved in AI to participate in international discussions.

This report is the original English version of a special issue published by the Japanese Society for Artificial Intelligence (JSAI), Vol.36, No.2 (March. 2021).

---

[1] https://ethicsinaction.ieee.org/

[2] http://www.ai-for-sdgs.academy/ai4sdgs-cooperation-network

[3] http://ai-elsi.org/archives/583

[4] https://www.partnershiponai.org/partners/

[5] http://www.aiandsociety.org/, and the report of this symposium is published as a special feature in Artificial Intelligence, vol. 33, no. 2.

Arisa Ema

Arisa Ema is an Associate Professor at the University of Tokyo and Visiting Researcher at RIKEN Center for Advanced Intelligence Project in Japan and Research Support Advisor at AIST Department of Information Technology and Human Factors. She is a researcher in Science and Technology Studies (STS), and her primary interest is to investigate the benefits and risks of artificial intelligence by organizing an interdisciplinary research group. She is a co-founder of Acceptable Intelligence with Responsibility Study Group (AIR) established in 2014, which seeks to address emerging issues and relationships between artificial intelligence and society.

She is a member of the Ethics Committee of the Japanese Society for Artificial Intelligence (JSAI), which released the JSAI Ethical Guidelines in 2017. She is also a board member of the Japan Deep Learning Association (JDLA). She was also a member of the Council for Social Principles of Human-centric AI, The Cabinet Office, which released "Social Principles of Human-Centric AI" in 2019. She obtained a Ph.D. from the University of Tokyo and previously held a position as Assistant Professor at the Hakubi Center for Advanced Research, Kyoto University.

Japan's Role in Multilateralism on AI Governance

Danit Gal Leverhulme Centre for the Future of Intelligence, the University of Cambridge

Introduction

With multilateral efforts on AI governance taking center stage globally, Japan is emerging as a leading actor. Home to a uniquely AI-embracing society, the country must strike a balance between distinct domestic views and uses of AI and mounting pressures to align with other AI world leaders in search of consensus-based AI governance. Taking this challenging task head on, Japan appears to be doing a good job of juggling domestic interests and international agendas. However, considering Japan's historically rich AI culture and the fact that multilateral efforts on AI governance are still in their infancy, this juggling act is far from concluded. This article examines local- and international-facing Japanese approaches towards AI governance to try and understand how the two might align and diverge in the present and future.

AI Governance in Japan

The governance of AI, in Japan and otherwise, consists of a multitude of factors shaping how the technology is perceived, designed, utilized, and regulated. Japan, in particular, has a unique combination of such factors due to its long and rich relationship with technology, AI being no exception. Alone, Japan's religious, cultural, social, political, and economic ties to AI cannot account for the country's prominent acceptance and adoption. Together, these factors paint a striking picture of unprecedented integration between human beings and intelligent technologies. Such integration is supported by non-negligible religious, cultural, social, and political beliefs and practices, adding a unique dimension to the common infrastructural integration of AI systems in Japan, largely motivated by economic realities and ambitions.

Religiously speaking, techno-animism is deeply rooted in Japan's two main religions: Buddhism and Shinto. Japan's Buddhist tradition emphasizes human-environment co-existence by acknowledging that everything has the nature of Buddha and thus the potential for enlightenment. This has proven true for robots as well, with android Kannon Mindar created to deliver Buddhist sermons, [1] SoftBank Robotics' Pepper automating Buddhist funeral rites, [2] and defunct Aibo robot dogs being the subject of Buddhist funeral rites themselves. [3] This is being extended to AI, increasingly deployed to augment these robots and further advance religious integration in Japan.

Shinto beliefs also have strong techno-animistic foundations, but of a somewhat different nature. In the case of Shinto techno-animism, the spiritual essence of gods and heroes may dwell in animate and inanimate objects, giving them human-like characteristics, spiritual importance and, at times, mystical properties. The embodiment of gods in physical objects and their animation supports the integration of natural and human-made objects, like technology, into human society in a harmonious manner. This is exemplified by the extension of Shinto rites to automobiles and robots, with Kiyomori, a humanoid robot co-developed by Tmsuk and Waseda University, taken to Munakata Taisha Shinto Shrine to pray alongside shrine maidens for the robot's safety and industry success. [4] This serves as yet another example of AI's ongoing religious integration in Japan, through Shinto's techno-animistic traditions.

Culturally speaking, Japan is a key influencer and exporter of popular culture portraying intelligent robots as guardians, heroes, friends, family members, and potential love interests. Popular shows like Astro Boy (鉄腕アトム) and Doraemon (ドラえもん), who even became Japan's anime ambassador, [5]

are clear early examples. Joining them are numerous other Japanese comics and animated shows that have significantly enriched this narrative through the years. [6] These animated shows include, but are not limited to: 絶 対彼氏 or Absolute Boyfriend (2008), which was readapted in multiple Asian countries; 僕 の彼女はサイボーグ or Cyborg She (2008); キュート or Q10 (2010); イヴの時間 or Time of Eve (2010); ちょびっツ or Chobits (2011); 安堂ロイド~ A.I. knows LOVE? or Ando Lloyd—A.I. Knows Love? (2013). [7] This popular cultural heritage is reflected in the inspiration and motivation of many Japanese AI researchers and developers creating AI-enabled artifacts that shape and are shaped by popular culture.

The acceptance and pursuit of AI-enabled artifacts as potential heroes, guardians, friends, family members, and even love interests underscore the considerable degree of AI's cultural integration in Japan. This might explain why over 3,700 Japanese individuals have been pursuing marriage certificates for their union with an AI-enabled holographic virtual assistant character named Miku, modeled after the virtual celebrity Hatsune Miku (初音ミク). [8] The pursuit of marriage between humans and an AI-enabled wife-virtual assistant hybrid represents a pinnacle in AI's cultural integration in Japan. The transition from a popular virtual celebrity to a romantic AI-enabled partner is clearly inspired by and made possible due to Japan's popular cultural. It also, however, holds considerable ramifications for AI's social integration in Japan as the technology assumes an evolving role in an increasingly lonely society. [9]

Socially speaking, many Japanese citizens have embraced AI as a social entity. This is evident in the celebrity status gained by AI-powered chatbot Rinna (りんな). Modeled after a schoolgirl and developed by Microsoft, Rinna is deployed on Japan's popular LINE messaging app. [10] Unlike virtual idols that gained popularity due to their artistic performances, Rinna gained popularity due to her AI-powered conversational capabilities. This demonstrates a closer degree of social acceptance and integration as an entity to be readily engaged rather than just viewed and admired. Offering another remarkable example of such social integration is AI-powered chatbot Shibuya Mirai, developed by Microsoft and deployed on LINE for Tokyo's Shibuya Ward municipality. Shibuya Mirai is modeled after a 7-year-old boy, designed to help neighborhood residents access local services and connect with local authorities. The chatbot is the first of its kind to have received local residency status, [11] emphasizing its unique, government-endorsed, social integration into the ward's day-to-day functions.

Further institutionalizing AI's social integration in Japan is the Japanese Society for AI (JSAI)'s 2017 Ethical Guidelines. The guidelines delineate expected behavior on the part of AI researchers and developers, aimed at benefiting society. Most uniquely, the last article (number 9), extends such responsibility to AI-systems themselves. It reads: "Abidance of ethics guidelines by AI) AI must abide by the policies described above in the same manner as the members of the JSAI in order to become a member or a quasi-member of society." [12] JSAI's guidelines give AI a chance to become part of society if it abides by them, guiding and normalizing such social integration.

Politically speaking, Japan has been among the earliest countries to examine the societal and ethical implications of technology, issuing white papers, guidelines, principles, and policy recommendations. [13] It did so, at least in part, because the country also produced one of the most advanced visions of a technology-enabled society, Society 5.0. This vision, introduced in Japan's 5th Science and Technology Basic Plan on January 22nd, 2016, portrays a future in which Japan's society co-exists and co-evolves with robots, AI, and other key technologies. This vision embeds such technologies in numerous core infrastructures to support rapid response to various human needs, with the goal of anticipating such needs and responding to them before they arise. [14] Over three years after its publication, the Society 5.0 vision continues to guide governmental thinking on the role of technologies like AI and robotics in Japan. This is evident in a 2019 series of government-issued videos explaining how this vision is being realized to enhance human ability and support the development of Japan's society. [15,16,17]

Society 5.0 has shaped subsequent government documents like the Cabinet Office's Social Principles of Human-Centric AI, providing guidelines for the creation of an "AI-based human living environment" for a "society premised on AI". This suggests strong political support for AI integration in Japan,

extending beyond infrastructural integration. In contradiction with the abovementioned religious, cultural, and social integration, however, the principles also warn against over-reliance on AI and the potential loss of human dignity that might come about if the warning is ignored. [18] This cautionary note and the growing emphasis on human-centric AI were described by Takehiro Ohya, professor at Keio University, as positions intended to foster stronger alignment with Western understandings of the role and place on AI in society. [19] This marks a shift in how Japan thinks about domestic versus multilateral efforts on AI and informs the nation's search of international consensus on AI governance.

Economically speaking, the creation of a technology-enabled Japanese society is an unprecedented economic boon motivating AI's infrastructural integration. As of October 2019, Japan's population decreased, for the ninth consecutive year, by 0.22%, marking the largest decrease margin documented to date. [20] With a super-aging and shrinking population, Japan stands to greatly benefit from the widespread deployment of robots and AI-enabled systems to complement a dwindling workforce. This would also lessen the pressing need for labor immigration, a contentious political topic in the country. [21] Indeed, Japan now serves as an internationally acclaimed model of AI and robots' integration in healthcare, retail, education, transportation, and other public service infrastructures. In 2018, Japan was recognized as the world's fourth most automated economy. But these achievements are also accompanied by concerns that automation will eventually lead to widespread job loss and cause greater inequality. [22] While the prospects of AI's infrastructural integration appeal to many in economic terms, the social cost of such automation, particularly of the intelligent kind, remains unclear.

This concern is further fueled by the fact that while Japan constitutes a large consumer market for AI, with an anticipated US$4.8 billion-worth Deep Learning market by 2025, [23] the country only accounted for around 2% of published AI research papers as of 2019. To fully benefit from the surging demand for AI's infrastructural integration, Japan will need to balance its capitalization on a world-leading hardware industry with speeding up the research and development of software in general, and AI-related software in particular. [24] Despite being the most commonly endorsed type, AI's infrastructural integration in Japan is not without its concerns and cannot be explained while detached from religious, cultural, social, and political motivations for integration.

Together, these factors illuminate Japan's unique relationship with AI, set to define how the country governs the integration of AI in many of its national facets. Much remains to be seen as to which AI-related policies, national standards, and testing tools Japan will develop and use. However, the high degree of economically-motivated infrastructural integration and the significant degree of religious, cultural, social, and political integration hint at a uniquely-Japanese approach to AI governance, at least domestically.

Japan in Global AI Governance

In the forementioned search for international consensus on AI governance, Japan plays a very active role. Japan was among the members of the OECD's Council on AI, publishing the OECD's Recommendations on AI, adopted by members on May 21st, 2019. This work "provides the first intergovernmental standard for AI policies," [25] and paved the path for the creation of the AI Group of experts at the OECD (AIGO), of which Japan is an active member. [26] The OECD's recommendations served as the foundation for the G20's AI Principles, adopted on June 2019 in Tsukuba City, Ibaraki Prefecture, Japan. [27] The principles offer guidelines for the responsible stewardship of trustworthy AI, national policies, and international cooperation on trustworthy AI. Japan played a key role in helping craft these principles and presenting them to other G20 members, playing an instrumental role in another successful multilateral effort on AI governance.

Japan is also a prominent member of the Global Partnership on AI (GPAI), launched in June 2020. GPAI was established to "provide a mechanism for sharing multidisciplinary research and identifying key issues among AI practitioners, with the objective of facilitating international collaboration, reducing duplication, acting as a global reference point for specific AI issues, and ultimately promoting trust in and the adoption of trustworthy AI." [28] At GPAI, Japan co-chairs the working group on the future of work, [29] sits on the steering committee, [30] and has expert members participating in other working groups.

While not a European country, Japan holds an observer status at the Council of Europe's Ad hoc Committee on AI (CAHAI). [31] CAHAI was established by the European Council's Committee of Ministers in 2019 to "examine the feasibility and potential elements on the basis of broad multi-stakeholder consultations, of a legal framework for the development, design, and application of artificial intelligence, based on the Council of Europe's standards on human rights, democracy, and the rule of law." [32] Japan also contributes expert participation and has funded CAHAI's publication "Towards Regulation of AI Systems." [33]

Including entities beyond governments, Japan participates in additional multilateral efforts like UNESCO's efforts to "create the first global standard-setting instrument on ethics of AI." [34] Japan reiterated its commitment to cooperation with UNESCO back in 2019, specifically naming AI as one of the core areas of collaboration in support of Africa's development. [35] An additional non-governmental multilateral effort on AI governance in which Japanese entities actively participate is the Partnership on AI. Japanese members include SoftBank, Sony, and the University of Tokyo's Next Generation Artificial Intelligence Research Center (AI Center). [36]

In 2019, Japan entered into a trilateral initiative, the French-Japanese-German Research Projects on AI. Co-led by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), the French National Research Agency (ANR, France), and the Japan Science and Technology Agency (JST, Japan), this trilateral initiative aims "to present the direction of future digital economy and society through technical progress in AI research to strengthen trust, transparency and fairness as well as improving performance and investigating AI methods." [37]

In 2020 alone, Japan has ramped up bilateral cooperation and released joint statements on strengthening cooperation relating to AI, among other technologies, with the United States of America, [38] the European Union, [39] and India. [40] This joins a slew of other existing bilateral agreements on AI cooperation with many expected to join the ranks in the years to come as Japan continues to bolster its international participation in and co-leadership of multilateral efforts on AI governance.

The Role of Japan in Multilateralism on AI Governance

While very actively charting the path towards international consensus on AI governance, Japan will also need to navigate the aforementioned uniqueness of its domestic market. On the far end of the spectrum, this uniqueness has previously led to Japan's market isolation, known as the Galápagos Syndrome. Japan coined this term after experiencing market isolation when rushing to design and deploy its home-made 3G Telecommunication network, only to find other countries aligned on a different network standard later on. This reverberated across the country, with Japanese mobile phones rapidly becoming incompatible and unmarketable abroad while foreign mobile phones suffered incompatibility issues at home. [41] This vivid example of market isolationism as an unintended consequence of domestic-facing innovation serves as a cautionary tale of the potential pitfalls one can expect in an overly domestic-facing Japanese AI ecosystem.

Unlike many other technologies, including robotics, this does not appear to be the case with AI. The aforementioned search for international consensus on AI governance shows that Japan has learned from past mistakes and is actively pursuing international cooperation and alignment. In particular, Japan seems to be placing a strong emphasis on trustworthy AI development and regulation. In the context of AI, trust is a complex concept encompassing technical, ethical, societal, and regulatory measures. [42] Thus far, Japan's pragmatic stance on trustworthy AI has yet to breed any international misalignment with its demonstrated religious, cultural, social, and political integration at home. But policy concerns of over-dependence on and demonstrated romantic attachment to the technology may end up rocking the domestic boat, considering the non-negligible degree of religious, cultural, social, and, albeit contradictory, political integration discussed in this article. This illustrates the delicate balance Japan maintains between multilateral efforts to achieve international consensus on AI governance and a demonstrated national appetite for the integration of AI. It remains to be seen how and whether this balance can be sustained.

It should also, however, be noted that over-alignment with a mostly Western-based consensus that differs from Japan's

views on and applications of AI risks marginalizing the country's unique heritage. Japan's religious, cultural, societal, and political integration of AI add further dimensions that are critical for the realization of the government's Society 5.0 vision of co-evolution and co-existence with robots, AI, and other key technologies. This charts the path towards a possible clash if international actors demand alignment in Japan's domestic market to maintain its international position. Criticism over the unconventional design and applications of AI in Japan and the resulting ethical concerns it amplifies and creates has been present for years now. [43] As these discussions grow in importance in Japan and abroad, some AI-enabled artifacts designed for social integration will become increasingly contested. If true, this will likely complicate Japan's participation in multilateral efforts on AI governance, particularly vis-à-vis Western countries.

For now, Japan's domestic approach towards AI and that of other, mostly Western countries are not mutually exclusive. This is evident in the leading role Japan plays in multilateral efforts on AI governance abroad and its continued push towards the multi-dimensional integration of AI at home. As such, Japan serves as an important and fascinating case study of how countries may be able to balance a singular local AI culture with successful participation in multilateral AI governance efforts. As more countries join these still largely exclusive multilateral efforts, Japan's participation model will likely increase in both visibility and importance.

Conclusions

Ultimately, the role Japan plays in multilateral efforts on AI governance will be shaped by many factors, present at home and abroad. Cognizant of its isolationist past, historically extending far beyond the above discussed Galápagos Syndrome instance, Japan appears to be succeeding in the difficult task of striking a balance between distinct local- and global-facing AI governance efforts. Many twists and turns remain to be discovered and overcome as the country marches towards an AI-enabled future society with unprecedented levels of AI integration and a leading global position on AI governance. For now, this balancing act has proven viable. Even if it falters, it will still serve as an imporant model for the further study of nations' participation in multilateral efforts on AI governance.

[1] The Japan Times. 2019. Kannon Bodhisattva Robot Unveiled at Kyoto Temple to Share Buddha's Religious Teachings. Retrieved from https://www.japantimes.co.jp/news/2019/02/23/business/tech/robotic-kannon-unveiled-kyoto-temple/

[2] Reuters. 2017. In Japan, Robot-For-Hire Programed to Perform Buddhist Funeral Rites. Retrieved from https://www.reuters.com/article/us-japan-robotpriest-idUSKCN1B3133

[3] Eiraku Maiko. 2019. Backstories: A Funeral for Dead Robot Dogs. Retrieved from https://www3.nhk.or.jp/nhkworld/en/news/backstories/346/

[4] Naho Kitano. 2007. Animism, Rinri, Modernization; the Base of Japanese Robotics. Retrieved from http://www.roboethics.org/icra2007/contributions/KITANO%20Animism%20Rinri%20Modernization%20the%20Base%20of%20Japanese%20Robo.pdf

[5] Japan Today. 2008. Doraemon Named 'Anime Ambassador'. Retrieved from https://japantoday.com/category/entertainment/doraemon-named-anime-ambassador

[6] Dorothy Florence Holland-Minkley. 2010. God in the Machine: Perceptions and Portrayals of Mechanical Kami in Japanese Anime. Masters Thesis. University of Pittsburgh, USA. (Unpublished)

[7] Danit Gal. 2020. Perspectives and Approaches in AI Ethics: East Asia. In Markus Dubber, Frank Pasquale, and Sunit Das (Eds.). Oxford Handbook of Ethics of Artificial Intelligence. Oxford University Press.

[8] Stephanie Hegarty. 2019. Why I 'married' a Cartoon Character. Retrieved from https://www.bbc.com/news/stories-49343280

[9] Michael Hoffman. 2018. Japan Struggles to Keep Loneliness at Arm's Length. Retrieved from https://www.japantimes.co.jp/news/2018/11/10/national/media-national/japan-struggles-keep-loneliness-arms-length/

[10] Microsoft Asia News Center. 2018. Rinna The AI Social Chatbot Goes Out and About In Japan's Countryside. Retrieved from https://news.microsoft.com/apac/2018/09/18/rinna-the-ai-social-chatbot-goes-out-and-about-in-japans-countryside/

[11] Microsoft Stories Asia. 2017. AI In Japan: Boy Bot's Big Honor. Retrieved From https://news.microsoft.com/apac/2017/11/20/ai-japan-boy-bots-big-honor/

[12] The Japanese Society for Artificial Intelligence Ethical Guidelines. 2017. Retrieved from http://www.ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf

[13] Arisa Ema. 2017. EADv2 Regional Reports on A/IS Ethics: JAPAN.

Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/eadv2_regional_report.pdf

[14] Government of Japan. 2016. The 5th Science and Technology Basic Plan. Retrieved from https://www8.cao.go.jp/cstp/english/basic/5thbasicplan.pdf

[15] Prime Minister's Office of Japan. 2019. Society 5.0: Highlights. Video. Retrieved from https://www.youtube.com/watch?v=S9JMuwvzz8g

[16] Prime Minister's Office of Japan. 2019. Society 5.0: Human Ability. Video. Retrieved from https://www.youtube.com/watch?v=odjuqbLJRMY

[17] 株式会社ダイヤサービスDAIYASERVICE Inc. 2020. 政府広報「Society5.0」. Video. Retrieved from https://www.youtube.com/watch?v=249hXyODqwY

[18] Cabinet Office Council on the Social Principles of Human-centric AI. 2019. Social Principles of Human-centric AI (Draft). Retrieved from https://www8.cao.go.jp/cstp/stmain/aisocialprinciples.pdf

[19] Danit Gal. 2020. Perspectives and Approaches in AI Ethics: East Asia. In Markus Dubber, Frank Pasquale, and Sunit Das (Eds.). Oxford Handbook of Ethics of Artificial Intelligence. Oxford University Press.

[20] Nippon. 2020. Japan's Population Falls for Ninth Straight Year. Retrieved from https://www.nippon.com/en/japan-data/h00705/

[21] Daniel Moss. 2017. Graying Japan Wants Automation, Not Immigration. Retrieved from https://www.japantimes.co.jp/opinion/2017/08/28/commentary/japan-commentary/graying-japan-wants-automation-not-immigration/

[22] Andrew Sharp. 2018. Japanese Fear Automation Will Take More Jobs Than Foreigners. Retrieved from https://asia.nikkei.com/Economy/Japanese-fear-automation-will-take-more-jobs-than-foreigners

[23] ReportLinker. 2020. Artificial Intelligence (AI) Market Worldwide is Projected to Grow By US$284.6 Billion. Retrieved from https://www.globenewswire.com/news-release/2020/05/22/2037692/0/en/Artificial-Intelligence-AI-market-worldwide-is-projected-to-grow-by-US-284-6-Billion.html

[24] Guillermo García. 2019. Artificial Intelligence in Japan: Industrial Cooperation and Business Opportunities for European Companies. Retrieved from https://www.eu-japan.eu/sites/default/files/publications/docs/artificial_intelligence_in_japan_-_guillermo_garcia_-_0705.pdf

[25] OECD. 2019. OECD Principles on AI. Retrieved from https://www.oecd.org/going-digital/ai/principles/

[26] OECD. 2020. List of Participants in the OECD Expert Group on AI (AIGO). Retrieved from http://www.oecd.org/going-digital/ai/oecd-aigo-membership-list.pdfwatch?v=S9JMuwvzz8g

[27] Ministry of Foreign Affairs of Japan. 2019. G20 Ministerial Statement on Trade and Digital Economy. Retrieved from https://www.mofa.go.jp/files/000486596.pdf

[28] The Global Partnership on Artificial Intelligence. 2020. About GPAI. Retrieved from https://www.gpai.ai/about/

[29] The Global Partnership on Artificial Intelligence. 2020. Working Group on the Future of Work. Retrieved from https://www.gpai.ai/projects/future-of-work/

[30] The Global Partnership on Artificial Intelligence. 2020. GPAI Steering Committee. Retrieved from https://www.gpai.ai/community/steering-committee/

[31] The Ad Hoc Committee on Artificial Intelligence (CAHAI). 2019. Agenda item 8.3 Admission of Observers to the CAHAI. Retrieved from https://rm.coe.int/cahai-2019-05-admission-of-observers-to-the-cahai/168098ad43

[32] Council of Europe. 2020. CAHAI - Ad hoc Committee on Artificial Intelligence. Retrieved from https://www.coe.int/en/web/artificial-intelligence/cahai

[33] The Ad Hoc Committee on Artificial Intelligence (CAHAI). 2019. Towards Regulation of AI Systems. Retrieved from https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a

[34] UNESCO. 2020. Artificial Intelligence with Human Values for Sustainable Development. Retrieved from https://en.unesco.org/artificial-intelligence

[35] UNESCO. 2019. Japan Renews Commitment to UNESCO Through Enhanced Cooperation on Artificial Intelligence and Cultural Diversity. Retrieved from https://en.unesco.org/news/japan-renews-commitment-unesco-through-enhanced-cooperation-artificial-intelligence-and

[36] The Partnership on AI. 2020. Meet the Partners. Retrieved from https://www.partnershiponai.org/partners/

[37] OECD. 2020. Trilateral French-Japanese-German Research Projects on Artificial Intelligence. Retrieved from https://www.oecd.ai/dashboards/policy-initiatives/2019-data-policyInitiatives-26493

[38] U.S. Department of State. 2020. Joint Press Statement on the 11th U.S.-Japan Policy Cooperation Dialogue on the Internet Economy. Retrieved from https://www.state.gov/joint-press-statement-on-the-11th-u-s-japan-policy-cooperation-dialogue-on-the-internet-economy/

[39] The European Commission. 2020. EU and Japan Step Up Cooperation in Science, Technology and Innovation. Retrieved from https://ec.europa.eu/info/news/eu-and-japan-step-cooperation-science-technology-and-innovation-2020-may-26_en

[40] The Hindu. 2020. India, Japan Ink Pact to Enhance Cooperation in ICT. Retrieved from https://www.thehindu.com/news/national/india-japan-ink-pact-to-enhance-cooperation-in-ict/article33581425.ece

[41] Hiroko Tabuchi. 2009. Why Japan's Cellphones Haven't Gone Global. Retrieved from https://www.nytimes.com/2009/07/20/technology/20cell.html?em

[42] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2020. Trustworthy Artificial Intelligence. Electronic Markets. DOI: https://doi.org/10.1007/s12525-020-00441-4

[43] Dongwoo Kim. 2020. Advancing AI Ethics in Japan: A Q&A with Dr. Arisa Ema, Professor at University of Tokyo. Retrieved from https://www.asiapacific.ca/publication/advancing-ai-ethics-japan-qa-dr-arisa-ema-professor

Danit Gal

Danit Gal is Associate Fellow at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge and Visiting Research Fellow at the S. Rajaratnam School of International Studies at the Nanyang Technological University. She serves as the vice chair of the P7009 IEEE standard on the Fail-Safe Design of Autonomous and Semi-Autonomous Systems, member of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems executive committee, Founding Editor and Editorial Board member of Springer's AI and Ethics journal, Executive Committee member of the AI4SDGs Cooperation Network at the Beijing Academy of AI, and Advisory Board member of UKRI's Trustworthy Autonomous Systems Verifiability Node.

Paving an Intentional Path Towards Inclusion in AI Development

Tina M. Park Partnership on AI

1 Introduction: Prioritizing Inclusion at the Partnership on AI

Technology holds the possibility of generating both positive and negative effects on the lives of human beings and the world around us. This could not be truer for Artificial Intelligence (AI) and Machine Learning (ML) systems in particular. We are witnessing first hand both the tremendous good enabled by algorithms as we battle the COVID-19 pandemic, like the use of learning-prediction models to identify already existing drugs that could be repurposed to treat COVID-19 [22], as well as their potential for widespread and long-lasting harm. For example, in the United States, Stanford University's Medical Center used an algorithm to determine who would receive the first wave of COVID-19 vaccines, resulting in the exclusion of nearly all 1,300 Resident Physicians working on the frontlines of the pandemic for the hospital [13]. As this global pandemic enters its second year, misunderstandings of what algorithms are, how they work, and how they are created may diminish already weakening trust in public health and vaccine management, as people worry their lives are in the hands of mysterious "black boxes."

The Partnership on AI (PAI), a non-profit organization based in San Francisco, CA, is working towards a future where Artificial Intelligence empowers humanity by contributing to a more just, equitable, and prosperous world. PAI does this by bringing together diverse voices across global sectors, disciplines, and demographics, creating a trusted forum where practitioners and others can share ideas and practices for Responsible AI.

With nearly 100 Partner organizations, including major global technology companies, research centers, and human rights organizations, PAI creates venues to tackle difficult questions about the social impact of AI and ML technologies through both dialogue and data-driven research. In addition to facilitating conversations between experts and leaders in industry, academia, and civil society, PAI conducts research to produce impactful, evidence-based guidance for Partners, and the technology industry more broadly, on how to navigate some of the most pressing concerns related to AI and society.

For example, there has been growing concern about the lack of diversity among technology workers, particularly highly paid engineers and management-level leaders [27]. In addition to reflecting racial and ethnic bias and discrimination in hiring in the technology industry, as well as other barriers to entry in the sciences, the lack of diversity in the AI field is worrisome as it may lead to significant racial and other biases encoded within algorithms [16, 27]. In partneship with DeepMind, PAI launched a diversity, equity, and inclusion (DEI) research study focused on the experiences of women and other minoritized individuals in AI in order to better understand why non-male, non-White employees are leaving the AI sector in disproportionately high numbers and to provide guidance on creating more inclusive environments for those working in AI [7].

Another important area of concern for PAI is the inclusion of diverse voices in the development and deployment of AI and machine learning systems. Our newest research project, Methods for Inclusion, uses a multidisciplinary approach to identify approaches and practices that can be implemented by AI/ML developers and researchers to expand the perspectives and needs considered in the creation of AI/ML technologies.

2 Why Make Artificial Intelligence More Inclusive?

Inclusion is an important tenet of AI/ML development for several reasons. The most obvious benefit of an inclusive approach is the ability to expand who is served by (and who purchases) any given product or service. In other words, there is a business case to be made for inclusion. Even the best products or services are not usable or relevant to everyone; thus, adaptations need to be made to accommodate potential users with different needs. For example, while most Sony Playstation users may find the controller comfortable and easy to use, someone who lacks full mobility and use of both hands is unlikely to play games using that same controller [15]. Adaptations to the controller, or even game play itself like using speech command instead of a physical controller, expands the possible pool of Sony Playstation users to a much wider audience [3].

Thinking about inclusion and exclusion can also serve as a catalyst for problem-solving and the creation of solutions that are better for everyone. An often-repeated example of the benefits of inclusive design in urban planning, for example, is the "curb cut" [4]. The curb cut is that familiar dip in the raised sidewalk that creates a gentle slope to meet the street. The ubiquity of curb cuts in the United States is largely due to the activism of wheelchair users and other people with disabilities. Initially implemented to allow people in wheelchairs the ability to transition smoothly from the sidewalk to the street without assistance, other people found these curb cuts to be extremely useful too. People pushing strollers or heavy carts, travelers with wheeled suitcases, and other able-bodied pedestrians found curb cuts to be very useful additions to their environment. A feature initially designed with a specific audience in mind people in wheelchairs turned out to be an improvement for many others.

More importantly, the inclusion of diverse perspectives, particularly those representing non-white racial-ethnic identities [8, 23], non-male gender identities [5, 10], and experiences of the disand differently-abled [14], is a means to mitigate some of the harm that AI and machine learning systems are shown to cause on already disadvantaged and oppressed communities. In other words, inclusive development, design, and deployment of AI/ML systems may prevent further social harm and help lessen existing social inequalities.

Developing AI/ML systems that are free of social harm is no easy task. For example, Twitter recently came under fire because the ML-based algorithm the company used to crop images on its platform favored white faces over those of Black-identified people [19, 20]. Relatedly, in 2019 several women AI researchers also flagged bias in Twitter's cropping, having identified numerous instances when the faces of women were cropped out of preview thumbnails, focusing instead on their chest [2].

Aware of systemic bias in its own AI/ML systems, Twitter actively attempts to test for gender and racial bias in its algorithms. Yet, despite conducting bias analyses on the ML-based cropping system [1, 12], the company failed to identify this issue until a user came across it, two years after the ML-based cropping system was implemented. As a company, Twitter is also known for intentionally trying to diversify its employee base [6]. While broadening the diversity of perspectives among its engineers is a useful first step in mitigating bias, in this case it was not enough to identify the problem before the image-cropping algorithm was deployed. It required active participation of a broad base of users (and those concerned with bias and discrimination). Although it took a concerted effort of concerned users to finally draw Twitter's attention to the issue, the company did take the feedback seriously and re-examined their algorithms [1]. The responsible creation and deployment of AI/ML systems requires the participation of users, as well as those otherwise impacted by the technology, to design, develop, test, and improve the technology and effectively mitigate any social harms that might result.

For these reasons, PAI has worked since its earliest days to find ways to seek out input from marginalized communities and stakeholders who are not traditionally consulted during the AI/ML development process.

3 Combating the Inclusion Illusion

Technology developers have long been thinking about how to address these barriers to inclusion in their work. Participatory design approaches used in technology

development have been around since the 1970s, relying on different stakeholder engagement practices such as interviews, focus groups, user surveys, and system evaluations. Applications from the field of User Experience (UX) research are an important way for companies to understand how someone uses, interacts with, and generally experiences their product or service. Whether beta testing a new product with a set of trusted users or conducting focus groups with potential customers to get a better idea of what users want and need in the latest iteration of a product, this collaboration between developer and user is key to producing and launching a successful AI/ML product (or any product).

At PAI, we first began exploring the idea of working with people outside of the "technical" sphere in partnership with the Tech Policy Lab at the University of Washington, a PAI Partner with extensive expertise in applying value-sensitive design approaches to technology policy [21, 28]. In 2019, PAI worked with the Tech Policy Lab to implement their Diverse Voices methodology within PAI's ABOUT ML project, an initiative focused on establishing documentation practices throughout the AI/ML lifecycle to provide greater transparency to the systems created. The aim was to explicitly solicit views and feedback from communities who are often the least likely to be consulted in the formation of machine learning system documentation practices that nonetheless impact them. The Diverse Voices consultants coordinated three experiential expert panels in Seattle, WA to review and comment on the first draft of the ABOUT ML report.

PAI learned a lot from the careful way the Tech Policy Lab team applied their research methodology towards the aim of greater inclusivity. Specifically, it underscored for us two crucial benefits of incorporating a wider array of perspectives in technology development:

1. It generates important and meaningful insights for tech policy documents, highlighting potential harm or unusability, as well as other uses that were not previously considered.
2. It leverages the expertise of groups that are historically excluded from the development and deployment of technology in mitigating future harm from the use of that technology.

However, it is important to acknowledge that participation is not the same thing as inclusion when it comes to technology development. As demonstrated in other sectors, participation can be used as a disingenuous means to extract labor without proper compensation or credit [25]. Participation may also be used as a way to legitimize the status quo by collecting input without incorporating it into final outcomes and by maintaining boundaries between who is essential versus nonessential in the decision-making process [17, 18]. For example, the Diverse Voices methodology is thoughtful and intentional about respecting the contributions made by experiential experts, or those with a depth of experience and insights gained through life and professional experience, rather than formal education or training. They emphasize the importance of compensating and valuing experiential experts who participated in the panel for both their time and insights. However, the inherent power dynamic between an organization's project leaders and the team implementing a methodology cannot guarantee that a final technology product will reflect the input given by non-technical stakeholders. Additionally, participation itself is not tied to any particular value commitments, other than the belief that more input will result in a better outcome (in this case, product or service). For this reason, metrics of participation often focus on how many people were involved in the development process as a proxy for inclusion. In other words, it is very possible to get lots of "participation" without actually being "inclusive."

Inclusion also requires acknowledging that exclusion exists – that not everyone who should participate in the technology development process is allowed or able to participate to the same degree. Exclusion can occur for many reasons, ranging from a lack of awareness around who else could be included and historic practices rooted in biases and prejudices to institutional policies that explicitly seek to keep certain people out of technology development. Exclusion can also arise, even with a diversity of employees or users, when the contributions of those traditionally with less authority or power are undervalued or otherwise dismissed. This attention to the power dynamics that privilege the needs and opinions of some groups over others is an important distinction between simple participation and full-fledged inclusion. To address these nuances at PAI, we think about inclusion as a form of participation that is specifically oriented towards achieving a sense of integration within a group or

institution. Within the framework of "diversity and inclusion," inclusion means creating an environment in which people of many different backgrounds, experiences, and expertise, are involved and empowered to make decisions within the group or organization.

Therefore, achieving a sense of community consultation and empowerment within deployed AI/ML systems requires more than simply soliciting feedback from stakeholders. It requires mapping out pathways through which stakeholders can actually directly engage in the decision-making process, ideally becoming one of many decision-makers and directly influencing the creation and/or deployment of the technology. This means expanding participation beyond whom we normally consider "experts" on AI/ML technology, and identifying how those who hold non-technical knowledge and experiences can become necessary contributors to the development of successful AI/ML systems. It also means cultivating trustworthy relationships between everyone involved so different insights and opinions, including ones that may run counter to existing assumptions, can be freely shared and incorporated into the broader pool of knowledge used to inform the development of a new AI/ML system.

4 Beyond Participation: Methods for Inclusion at PAI

PAI believes that working with communities to develop products and services early and throughout the development process providing multiple touchpoints to assess who is served and how helps AI/ML developers mitigate potential harm or negative impact. A truly inclusive approach can help developers build long-lasting, trusting relationships with the people they want to ultimately serve through their technology.

In order to deepen our understanding of this issue, we created the Methods for Inclusion Fellowship to commit time and resources to research and create materials for those in the AI/ML development community to better capture the nuances of participation and inclusion.

The Methods for Inclusion project is foremost attentive to dynamics of systemic power inequality which have historically resulted in the exclusion and neglect of certain communities and populations from the AI/ML development lifecycle and process. It also extends inclusion beyond the direct participation of individuals representing specific identities or experiences by considering how non-human inputs (e.g., training datasets) serve to include or exclude. This project is multidisciplinary in nature, learning from fields outside of computer science and technology that have grappled with questions of participation and inclusion for many decades. This includes fields like civic governance, education, planning and policy, public health/healthcare, and the social sciences. The project also takes insights and guidance from community organizing, which can cover many topics and disciplines.

Through Methods for Inclusion, we are broadening the aperture in recognition of the other scholarship that exists on the topic of inclusion in various domains. The project builds on lessons we've learned from our valued Partner, the Tech Policy Lab, and the existing work of scholars, practitioners, and most importantly, advocates, who have, for years, tried to open up AI/ML development to people outside of the close circle of engineers and developers.

Throughout 2021, the Methods for Inclusion project will work to:

- identify a range of participatory practices from different fields that could be adapted for use by AI researchers, designers, and developers;
- better understand the challenges of incorporating inclusive methods into AI development, with a specific eye towards the different barriers and incentives facing AI developers, on the one hand, and members of impacted communities on the other; and
- create real-life case study resources that outline attempted participatory methods, the challenges faced by each, and the improvements experienced by companies as a result.

Ultimately, through Methods for Inclusion, we hope to place AI developers and community members around the world who are invested in avoiding potential harm resulting from AI/ML systems into direct conversation with one another.

5 Inclusion as a Global Issue

It is easy to characterize inclusion and exclusion, particularly as it relates to racial bias and discrimination, as solely an issue unique to the United States. However, to believe this would be to deny the histories of migration, immigration, and colonization around the world that have resulted in non-homogeneous societies in every nation. Furthermore, exclusion occurs based on multiple, and often time overlapping, social dimensions such as gender, caste, or social class [9, 11]. Unfortunately, this means that no society is immune to exclusion, bias, and discrimination.

It is also important to recognize that exclusionary values and practices can and do travel. Anti-Black racism is not exclusive to Americans; the belief that Black people are somehow less capable or qualified extends beyond the borders of the U.S., affecting how Black people are treated wherever they may be. Ideas, and the everyday practices and behaviors that emerge from those ideas, circulate globally and embed themselves in organizational contexts far from their site of origin.

For example, Silicon Valley technology companies are currently facing a different kind of diversity issue: the issue of caste. A lawsuit has been brought against U.S-based technology companies for discrimination based on the Indian caste system [26]. Engineers who identify as Dalit, the lowest-ranked caste within India's social hierarchy, allege they experience difficulty getting hired for roles based outside of India because of caste-enforcing practices brought into non-Indian organizations by higher-caste Indian interviewers and hiring managers.

This highlights the importance of considering the presence of exclusion within the AI/ML development process, not only within the local context of one's own city, region, or nation, but throughout the various manifestations of bias and discrimination drawn globally. This is also important because technology itself is mobile. Technology developed in Japan may be used in the U.S., Nairobi, or Brazil. To be an ethical and responsible AI/ML researcher and developer is to recognize that the abuse of technology to deepen social inequality may happen far from where the technology was originally developed. It is important to initiate conversations around how technology not only helps or hinders social inequality from manifesting locally, but also how technology may be used and abused in different social contexts all around the world.

6 Conclusion: Paving Paths Towards Inclusion

Organizations can and should be proactive in their commitment to diversity and inclusion by auditing themselves and their research and development teams to assess the barriers that may stifle contributions made by traditionally excluded communities, such as women and gender non-conforming people, people with disabilities, and racial and ethnic minorities. Who is harmed by exclusion is not fixed and thus identifying those voices may vary from project to project. Working towards inclusion will require careful consideration of the organization itself, the local context (where the development is taking place), as well as the global context (where the technology may be deployed). Intentional pathways should be created so those who are excluded can meaningfully contribute to the design and development of AI technologies. Projects that focus on the social impact of AI technology should be supported and sponsored.

Ultimately, creating many more opportunities for AI/ML developers to learn about and critically examine social discrimination and bias is an important first step in producing responsible and inclusive AI.

Currently, their day-to-day job requirements make it possible for AI/ML developers to create products with wide-reaching global impact that will only compound over time. Thus, it is crucial to equip these technology developers with historical and socio-technical literacy so that they can begin to ask critical questions of the impact of their work and to seek out experts for deeper discussions. Moreover, this responsibility cannot rest solely with the individual employee, but rather must be incorporated through organizational processes (such as oversight, auditing, promotion, etc.). By embedding these practices into the overall organization's functioning, it makes it possible to root out biases and discrimination as a part of day-to-day practice [24]. It also supports individuals to act upon their ethical impulses, whether through

whistleblower protections, transparent and responsive decision-making processes, rewards for stopping the release of problematic features, or other mechanisms.

More active and regular conversations about the experiences of women and girls, people who identify as lesbian, gay, bisexual, transgender, or queer (LGBTQ), people with disabilities, or ethnic minorities, led by those with first-hand experience, can only improve the technology that is created. Thoughtful engagement requires being receptive to challenges to your status quo, including accepting that bias, discrimination, and social inequality exist and that everyone, even unintentionally, contribute to the maintenance of these divisions. By being mindful of who is excluded and the extent to which they are excluded from the development, use, and enjoyment of AI technologies, we can actively work towards greater inclusion.

7 Acknowledgments

Many thanks to Kemi Bello and Jingying Yang of Partnership on AI for their editorial and content support.

[1] Parag Agrawal and Dantley Davis. 2020. Transparency around image cropping and changes to come. https://blog.twitter.com/official/en_ us/topics/product/2020/transparency-image-cropping.html

[2] Anima Anandkumar. 2020. I had tweeted in 2019 about @ Twitter cropping #womeninAI headless while cropping men correcting. When I raised this, many men in field accused me of making up a non- existent issue just to gain attention. Sadly #AI #bias is not yet fixed. https://twitter.com/AnimaAnandkumar/status/1307465594304974848

[3] Jason M. Bailey. 2019. Adaptive Video Game Controllers Open Worlds for Gamers with Disabilities. The New York Times (Feb. 2019). https://www.nytimes.com/2019/02/20/business/video-game- controllers-disabilities.html

[4] Angela Glover Blackwell. 2017. The Curb-Cut Effect. Stanford Social Innovation Review (2017). https://ssir.org/articles/entry/the_curb_cut_ effect

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520 [cs, stat] (July 2016). http://arxiv.org/abs/1607.06520 arXiv: 1607.06520.

[6] Delana Brand. 2020. Inclusion & Diversity Report March 2020. https://blog.twitter.com/en_us/topics/company/2020/Inclusion- and-Diversity-Report-March-2020.html

[7] Jeff Brown and Alice Xiang. 2020. Beyond the Pipeline: Addressing Attrition as a Barrier to Diversity in AI. https://www.partnershiponai. org/diversityinai/

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Conference on Fairness, Accountability, and Transparency 81, 1 (2018), 1–15.

[9] Jane Coaston. 2019. The intersectionality wars. Vox (May 2019). https://www.vox.com/the-highlight/2019/5/20/18542843/intersectionality-conservatism-law-race-gender-discrimination

[10] Sasha Costanza-Chock. 2020. Design Justice: Community-Led Practices to Build the Worlds We Need. MIT Press, Cambridge, MA.

[11] Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. Stanford Law Review 43, 6 (July 1991), 1241. https://doi.org/10.2307/1229039

[12] Chaim Gartenberg. 2020. Twitter plans to change how image crop- ping works following concerns over racial bias. The Verge (Oct. 2020). https://www.theverge.com/2020/10/2/21498619/twitter-image- cropping-update-racial-bias-machine-learning

[13] Eileen Guo and Karen Hao. 2020. This is the Stanford vaccine al- gorithm that left out frontline doctors. MIT Technology Review (Dec. 2020). https://www.technologyreview.com/2020/12/21/1015303/ stanford-vaccine-algorithm/

[14] Aimi Hamraie and Kelly Fritsch. 2019. Crip Technoscience Manifesto. Catalyst: Feminism, Theory, Technoscience 5, 1 (April 2019), 1–33. https://doi.org/10.28968/cftt.v5i1.29607

[15] Kat Holmes. 2018. Mismatch: How Inclusion Shapes Design. The MIT Press, Cambridge, MA.

[16] Ayanna Howard and Charles Isbell. 2020. Diversity in AI: The Invisible Men and Women. MIT Sloan Management Review (Sept. 2020). https://sloanreview.mit.edu/article/diversity-in-ai-the- invisible-men-and-women/

[17] MSI Integrity. 2020. Not Fit-for-Purpose: The Grand Experiment of Multi-Stakeholder Initiatives in Corporate Accountability, Human Rights, and Global Governance. Technical Report. MSI Integrity.

[18] Natasha Iskander. 2018. Design Thinking Is Fundamentally Conser- vative and Preserves the Status Quo. Harvard Business Review (Sept. 2018). https://hbr.org/2018/09/design-thinking-is-fundamentally- conservative-and-preserves-the-status-quo

[19] Kim Lyons. 2020. Twitter is looking into why its photo preview appears to favor white faces over Black faces. The Verge (Sept.

2020). https://www.theverge.com/2020/9/20/21447998/twitter-photo- preview-white-black-faces

[20] Colin (colinmadland) Madland. 2020. A faculty member has been asking how to stop Zoom from removing his head when he uses a virtual background. We suggested the usual plain background, good lighting etc, but it didn't work. I was in a meeting with him today when I realized why it was happening. https://twitter.com/colinmadland/ status/1307111816250748933

[21] Lassana Magassa, Meg Young, and Batya Friedman. 2017. Diverse Voices: A How-To Guide Facilitating Inclusiveness in Tech Policy. Tech- nical Report. Tech Policy Lab, University of Washington, Seattle, WA.

[22] Sweta Mohanty, Md Harun Al Rashid, Mayank Mridul, Chandana Mohanty, and Swati Swayamsiddha. 2020. Application of Artificial Intelligence in COVID-19 drug repurposing. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14, 5 (Sept. 2020), 1027–1031. https://doi.org/10.1016/j.dsx.2020.06.068

[23] Safiya Umoja Noble. 2018. Algorithms of Oppression: How Search En- gines Reinforce Racism. New York University Press, New York, NY.

[24] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2020. Where Responsible AI meets Reality: Practi- tioner Perspectives on Enablers for Shifting Organizational Practices. https://arxiv.org/abs/2006.12358

[25] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a Design Fix for Machine Learning. In Proceedings of the 37th International Conference on Machine Learning, Vol. 119. PMLR, Vienna, Austria.

[26] Nitasha Tiku. 2020. India's engineers have thrived in Silicon Valley. So has its caste system. The Washington Post (Oct. 2020). https://www.washingtonpost.com/technology/2020/10/27/indian-caste-bias-silicon-valley/

[27] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating Systems: Gender, Race, and Power in AI. Technical Report. AI Now Institute, New York, NY. https://ainowinstitute.org/discriminatingsystems.html

[28] Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. Ethics and Information Technology 21, 2 (2019), 89–103. https://doi.org/10.1007/s10676-019-09497-z

Tina M. Park

Tina M. Park, Ph.D. Research Fellow at the Partnership on AI (PAI). Tina leads the Methods for Inclusion project at PAI, a project dedicated to developing evidence-based methodolo- gies to incorporate a more diverse range of stakeholders in the design and development of artificial intelligence. More information about can be found at www.partnershiponai.org. Tina earned her Ph.D. in Sociology at Brown University and her Master's in Urban Planning at New York University.

AI for Good and its Global Governance

Yolanda Lannquist, Adriana Bora, Niki Iliadis, The Future Society

Introduction

Artificial Intelligence (AI) has the capacity to unlock enormous opportunities, addressing major global challenges and achieving progress towards the UN Sustainable Development Goals (SDGs) – including breakthroughs in healthcare and education, access to goods and services, public service delivery, fairness at scale, and individual empowerment. However, at the same time, the same technologies pose risks and challenges. Amongst others, these include risks to the OECD AI Principles and require urgent action by policymakers, industry, citizens, and other actors.

Promoting the responsible development and use of AI means striking the right balance between capturing opportunities and mitigating risks. Since 2016, more than 150 responsible AI principles have been published internationally by governmental bodies, non-profits, academic institutions and companies. Most of these have taken the form of ethical guidelines or codes of conduct. Now, momentum is growing to put these principles into practice. Projects and frameworks to operationalize AI ethical principles and AI for social good, mechanisms to mitigate risks (e.g. bias & discrimination, cybersecurity, inequities), tools, certification methods, assessments and audit mechanisms have already sprung worldwide.

The rise of AI involves an unprecedented combination of complex dynamics, which poses challenges for multilateral efforts to govern its development and use. Global governance has a role to play in balancing the benefits and risks of deploying AI technologies, taking due care to ensure citizens are aware of their rights and protections. Meanwhile, it will have to balance different objectives, values, and incentives among stakeholders (policymakers, private sector, academia, civil society, etc.) and nation states. As a world leader in both economy and technology, Japan has a crucial role to play in leading the responsible adoption of AI. Building on its technological successes, Japanese industry, government, and academic leaders should pursue approaches to increase inclusion particularly of women in AI. Growing the participation of women will strengthen Japan's AI ecosystem and equip it to lead in responsible adoption of AI into the next decade.

AI for Sustainable Development Goals

While AI technologies pose risks and challenges to society, their potential for social good cannot be overlooked. With less than ten years to achieve the UN Sustainable Development Goals Agenda 2030, innovation and acceleration of technology adoption are needed. AI has the capacity to unlock enormous opportunities in societal, political, economic and cultural processes - including millions of lives saved by breakthroughs in healthcare, better personalization of products and services, easier access to public goods, and individual empowerment. It is important to balance the risks of "missed opportunities" to benefit from AI applications with the myriad of important ethical and safety risks.

Governance and ethical frameworks are required - ranging from data protection regulation to ethical guidelines - to capture the opportunities and mitigate the risks, within the context of a fast-evolving, possibly disruptive, and still uncertain technological landscape. Even in times of crisis, such as the COVID-19 pandemic, the need for an effective AI governance framework to uphold ethical standards and ensure alignment between companies and policymakers is necessary to mitigate the potential risks these technologies have to undermine our fundamental human rights and civil liberties.

There are numerous opportunities for AI applications to improve performance, efficiencies, and expand access to critical goods and services. AI has already been deployed to advance the work against some of the most pressing issues such as modern slavery, climate change, natural disaster or diseases. For example, with the use of real time data collected by satellites, mobile phones, and financial transaction technologies, AI can have a transformative impact on the management of Earth's natural resources and help achieve several SDGs. Satellite imagery, in particular, powered with AI capabilities can help predict, monitor, and evaluate environmental or agricultural changes. To have a complete and anticipatory view of natural disaster zones, and satellite imagery coupled with AI-based systems enables quick and effective decisions in times of crisis. Experiments suggest [1] that AI-powered systems can diagnose skin cancer with greater accuracy than human dermatologists. More recently, AI has been deployed to help measure, track and assess the scale of modern slavery across the globe. For instance, AI technologies can assist governments and companies to assess their risks of exploiting people under their supply chains and operations. With the use of natural language processing and machine learning, The Future Society's project AI Against Modern Slavery [2] analyzes and benchmarks the businesses' reports on modern slavery, assessing their compliance with targeted legislations.

Currently, AI plays a role in the responses to the COVID-19 pandemic. It can help governments conduct contact tracing, map the spread and forecast effects of different public health strategies. What is more, several AI-enabled solutions have been deployed to accelerate the discovery of vaccines and treatments. Yet, while many resources have shifted towards the response to the pandemic, the broader analysis of the socio-economical context demonstrates the need for an interdisciplinary approach to health outcomes at the national level. The multi-stakeholder initiative Collective and Augmented Intelligence Against COVID-19 (CAIAC) [3] brings key international actors to support informed policy decisions on pandemic responses.

International partners and diverse stakeholders are critical to accelerating AI adoption for social good. International support and partnerships on the use of AI for social good such as AI Commons [4], Global Data Access Framework [5], or Global Governance of AI Forum [6], are critical for coordinated progress. This cannot be done without the private sector, which can provide significant capacity and scale transformation. What is more, the deployment of AI systems is relatively in its early stages and citizens should also be proactively engaged to build a culture of trust. If people do not trust AI systems, these technologies will not reach their potential.

The public sector plays a monumental role, serving as an enabler, a facilitator, and a watchdog to ensure the process of AI implementation is ethical and benefits society broadlyl. As national governments mitigate the downside effects of AI and reconcile tensions between and across SDGs, it is important that they shape the right paths towards the technology's implementation as part of their national AI strategies. Therefore, governments should build the infrastructure for innovation, map the trajectories and (and sometimes trade-offs) among the SDGs and open data to support AI applications. They should ensure that talent will be educated and ready to participate in the race to achieve the SDGs Agenda 2030.

Technological advancements in AI are being enabled by greater digital connectivity, rapidly increasing amounts of data, advanced algorithms, and gains in computing and processing power. Yet, while AI development is rapid, adoption across markets is still at an early stage and much of its value is yet to be tapped. There is a window of opportunity to harness AI in developing contexts. The pace and magnitude of the digital revolution suggest that developing areas cannot afford to lag behind in leveraging AI.

Yet AI and digital technologies offer developing areas significant opportunities and risks in a 'more to gain, more to lose' paradigm. Countries can harness AI to address pressing social and economic problems. However, AI can also exacerbate important societal risks. Building on our existing work with the World Bank Digital Development Partnership analyzing international AI strategies, The Future Society is launching a new program to advise national AI strategies with a focus on the Global South. Our program builds frameworks of 'enabling' policies and practical implementation roadmaps for emerging and developing countries to build and harness

artificial intelligence for their development objectives while mitigating risks. In 2020-2021, The Government of Rwanda represented by the Ministry of ICT and Innovation (MINICT) and Rwanda Utilities Regulatory Authority (RURA) and GIZ FAIR Forward have engaged The Future Society to support the development of Rwanda's National AI Policy [7].

Industry vs. Policymakers: Diverging Ethical Priorities for AI

Despite the crucial role that AI has in the achievement of the SDGs Agenda 2030 and in fighting against the COVID-19 pandemic, even in times of crisis, the need for an effective governance framework and alignment between industry and policymakers is crucial to mitigate the potential risks these technologies have to undermine our fundamental human rights and civil liberties. Sound governance helps to bridge trust gaps and build a common understanding between policy and company leaders about the ethical risks raised by AI applications during the pandemic and beyond.

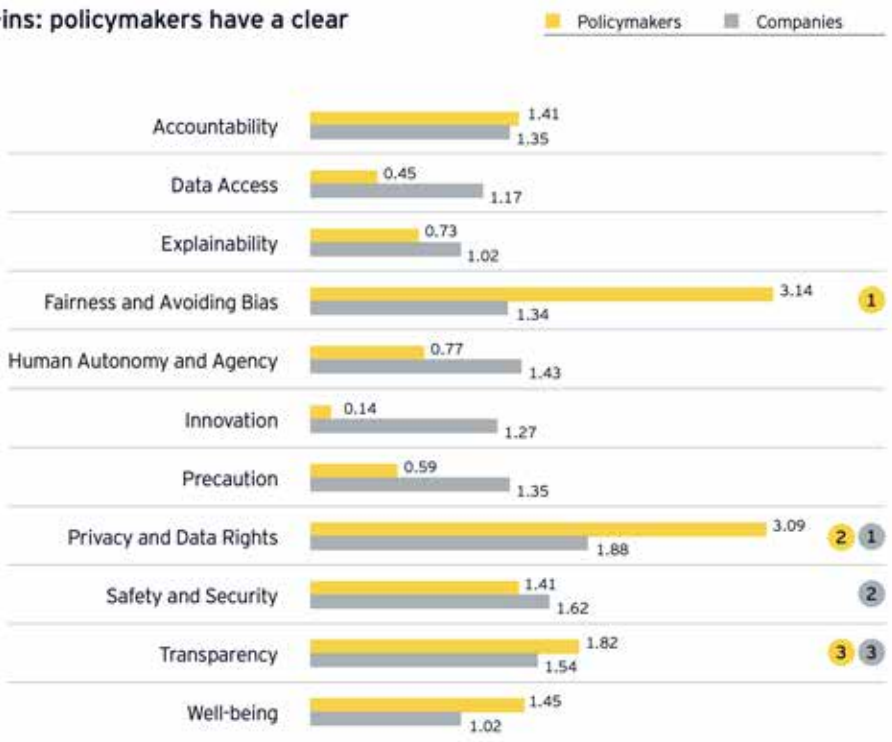
The Future Society and EYQ (EY's think tank) conducted a study to understand the ethical gaps between stakeholders that need to be addressed for the trustworthy adoption of AI across sectors. The global survey and report Bridging AI's trust gaps: Aligning policymakers and companies' [8]examines international AI ethical principles and the gaps in priorities for these principles across various AI applications in healthcare, aviation, law, retail, and financial services.

We asked policy and company leaders to identify the most important ethical principles when regulating a range of AI applications, and found divergent priorities between them across use cases. Ethical misalignments generally concentrate in four areas: fairness and avoiding bias, innovation, data access, and privacy and data rights.

The general trends of the study show that policymakers have a clear understanding and vision of the AI ethical risks, with a clear consensus on what needs to be prioritised in each context. On the other hand, companies have a much weaker consensus, who do not have clear collective priorities for the AI use cases surveyed.

The research shows that companies tend to select as priority

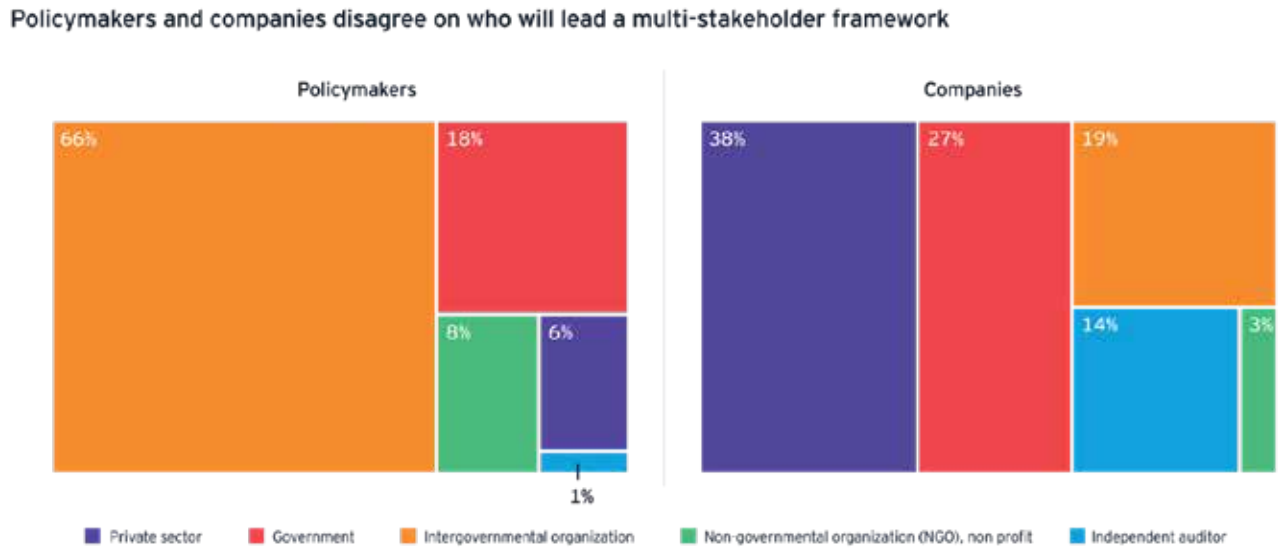
Exhibit 1. Facial recognition check-ins: policymakers have a clear vision of AI ethical principles



[Figure 1] Facial recognition check-ins:
policymakers have a clear vision of AI ethical principles, Cited from The Future Society and Ernst & Young, 2020

principals that are reflected in general regulations such as GDPR (e.g., privacy and cybersecurity) rather than on emerging issues that will become critical in the age of AI (e.g., explainability, fairness and non-discrimination). Companies' focus on the existing regulations could be explained by their incentives to maximize profits, whereas policymakers respond to broader public interest. Policymakers tend to have a larger vision and horizon and prioritise principles that have a social benefit even if they can be seen as less tangible, such as fairness, human autonomy and explainability.

What is more, the governance of AI requires alignment in the distribution of roles. Our study shows that companies and policymakers do not expect the same pathway for AI governance. While 38% of companies expect the private sector to lead a multi-stakeholder framework, only 6% of policymakers agree (instead, two-thirds of them think an intergovernmental organization is most likely to lead). Given the



[Figure 2]
Policymakers and companies disagrees on who will lead a multi-stakeholder framework, Cited from The Future Society and Ernst & Young, 2020

complex issues at play, both policymakers and companies expect that a multi-stakeholder approach will be needed to AI governance.

Global Governance and Coordination for AI

For nations and diverse stakeholders to harness AI opportunities at scale and mitigate their risks, there is a need for cohesive global cooperation and collaboration. International initiatives and platforms for governance can help to reconcile technical, ethical, commercial, legal and operational frameworks and protocols - to take the power of AI technologies and successfully make progress towards the achievement of the SDGs.

International organizations - such as UNESCO, the International Telecommunication Union, OECD, the Global Partnership on AI - and supranational government bodies - such as the European Union, the African Union, the Nordic-Baltic Region, the G20, and the G7 - have already started to move towards this direction, coordinating policies and pooling resources across countries to devise and implement AI strategies which will benefit all of humanity.

For example, several of these actors have developed principles or codes of conduct for how AI should be developed and deployed. One of the most commonly used sets of principles is that of the OECD, agreed upon by OECD members, including Japan, and non-members such as Argentina, Brazil, Colombia, Costa Rica, Peru, and Romania. The OECD AI Principles prioritize i) inclusive growth, sustainable development and well-being; ii) human-centred values and fairness; iii) transparency and explainability; iv) robustness, security and safety; and v) accountability. Although they are not legally binding, they set the ethical principles for how AI should be developed and deployed, and raise awareness for global coordination over AI.

Beyond principles, global coordination and governance can also take different forms. Mechanisms of hard governance are laws and regulations such as GDPR while mechanisms for soft governance that have emerged include standards from organizations like IEEE, certifications and training, public awareness campaigns, risk assessments, and audit mechanisms. Governance can also take the form of transnational coordination initiatives such as the recently launched Global Partnership for AI (GPAI). GPAI is an international and multistakeholder initiative to undertake applied AI projects and provide a cross-national mechanism for sharing multidisciplinary analysis, foresight and coordination - with the objective of facilitating international collaboration and synergies, and reducing duplication in the area of AI systems. The initiative was launched in June 2020 by Canada and France - along with Australia, Australia, Germany, India, Italy, Japan, Mexico, New Zealand, the Republic of Korea, Singapore, Slovenia, the United Kingdom, the United States and the European Union. They were joined by Brazil, the Netherlands, Poland and Spain in December 2020.

GPAI has established five working groups which have each launched research projects to identify coordinated approaches. In 2020, The Future Society led research for the Responsible AI and AI's Pandemic Response [9] working groups, while participating as an expert in the Data Governance Working Group, in order to understand opportunities for future action in the ecosystem.

A key conclusion from the review is that although the global landscape is moving towards a more coordinated approach, more needs to be done to reach a truly global "regime complex" which aligns the rise of AI with transnational fundamental rights. Global coordination and governance requires gathering information from dispersed sources, shining light on issues that are often unexamined, and encouraging knowledge-sharing and debates across disciplinary, sectoral, regional, and cultural divides.

Japan as a Leader in Global Governance of AI

Japan has played a leading role in the establishment of governance policies and norms for responsible AI development and adoption. Japan and a number of democratic countries share a vision for a human-centric development of AI grounded in ethical principles. Its role on the global stage was highlighted at the French-German-Japanese Symposium on Human-Centric Artificial Intelligence in November, 2020. Japan in particular has been a pioneer in the global governance of AI. For example, under the leadership of Mr. Yoichi Iida, Deputy Director General for G7 and G20 Relations, Chair of Committee on Digital Economy Policy at OECD, and Ministry of Internal Affairs and Communications, Japan has led activity that gave rise to the OECD AI Principles, now adopted by the G20 countries, and the OECD AI Policy observatory to implement those principles. Japan also participates in the GPAI.

Japan led in the early development of a national AI strategy, which has since inspired a proliferation of national AI strategies around the world. There are also a number of initiatives [10] at the national level to promote responsible AI research and development, including AI R&D Guidelines (The Conference toward AI Network Society), Social Principles of Human Centric AI (Integrated Innovation Strategy Promotion Council), and Governance Innovation: Redesigning Law and Architecture for Society 5.0 (Ministry of Economy, Trade and Industry) [11].

The Urgent Need for Women in AI

Leaders in Japanese government, industry, academia, and civil society should promote more women in the AI field. While Japan's AI ecosystem has many advantages, it shares the same shortage of AI talent faced by most if not all countries. One practical way to close the gap and support Japan's competitiveness in AI is to increase inclusion of women and other populations into the AI field.

Globally, women are underrepresented through the entire chain of AI research and development. More recent studies bring to light the worrying statistics showing that computer science as a whole is experiencing a historic low point for diversity. Female students are only 28% of those enrolled in information and communication technologies worldwide [12]. What is more, women still only make up a low 12% of the machine learning workforce and only a quarter of all

STEM workers are female, and even fewer are in positions of leadership.

Studies from the United States [13] show that women make up only 18% of computer science majors as of 2015 (down from 37% in 1984). Representing only 24% of the computer science workforce, women are also having salaries as 66% of those of their male counterparts. What is more, 50% of American women in technology drop by the age of 35 [14]. In companies such as Facebook and Google, women comprise only 15% and 10% respectively of AI research staff.

Women in AI are also under-represented in AI publications and conferences [15]. Only 18% of authors at leading AI conferences are women. While in 2019, women made up 23.2 percent of all computer science PhD students, only 10.6 percent of publications at ICLR 2020 had a female first author.

While Japan is leading in many aspects of AI development, according to the 2018 Women in Tech Index [16] it has a notable lack of women in the sector. With a large pay gap of 32%, Japan has a diversity crisis in the technology sectors and AI industry.

Women technology leaders such as Shelley McKinley Head of Technology & Corporate Responsibility at Microsoft explain [17] "As an industry and a society, we have a shared opportunity and responsibility to influence how technology, and specifically AI, accelerates our efforts to empower every person and organisation on the planet to achieve more. We must address the need to deploy technology in a responsible and inclusive way." In order to achieve this, we also need to address our inherent human biases and the lack of representation in who designs AI systems and algorithms. Without the input of women, technology is left vulnerable to many biases in design. Fair and robust technologies are dependent on a balanced and diverse workforce.

Female role models can provide support networks and create a sense of solidarity amongst women interested or already working in technology and AI. Platforms under which women can share their experiences and journeys in the technology industry can help shift gendered perceptions of the industry and generate change. Also, AI conferences should review processes and take measures to ensure all researchers get a fair chance to present their work. Initiatives such as Women in AI [18], Women in AI Ethics [19], AI4All [20], Elements of AI [21], European Women in Technology [22], Women of Silicon Valley [23], Women in Tech Africa and Women of Silicon Roundabout [24], etc. feature female speakers from the sector who inspire other women. Japan hosts one of the Ambassadors of the Women in AI [25], Eriko Toda, CEO at HappyCom.

Japan's failure to close the gender gap in technology and AI in particular, could have societal and economic consequences, which might affect its potential to reach its leading role in the AI global race. Without urgent sound changes in its gender policy, Japan will risk having a greater gender inequality while also diminishing the AI's economic and societal potential.

Concluding remarks

Embedded in the digital revolution, AI will play a factor in determining our societies for decades to come, accentuating and accelerating the dynamics of an old cycle in which technology and power reinforce one another. While AI has the potential to dramatically improve lives and make remarkable progress towards achievement of the SDGs, the socio-economic opportunities are inextricably connected with serious risks to the OECD AI Principles. Comprehensive national AI policies, AI ethical guidelines, and practical implementation plans fit for local contexts can serve as powerful roadmaps to achieve development objectives. These require broad stakeholder and public input and include national AI talent development - including women and minorities. Policymakers and companies must align on the key concerns for diverse AI use cases and technologies. With cross-national deployment and impacts, international collaboration and coordination is also needed in valid institutions such as the GPAI, OECD, Global Governance of AI Forum.

About The Future Society

The Future Society is an independent nonprofit 'think-and-do-tank' originally incubated at Harvard Kennedy School in 2014. The Future Society's mission is to advance the responsible adoption of AI and emerging technologies to

benefit humanity. The Future Society is in 'Best Think Tanks for Artificial Intelligence' by University of Pennsylvania Lauder Institute's Global Go To Think Tank Index. Find the list of our AI policy research, advisory services, seminars & summits, education projects, events, AI technical projects at https://thefuturesociety.org.

[1] https://www.sciencedaily.com/releases/2018/05/180528190839.htm

[2] http://thefuturesociety.org/2020/06/23/project-aims-artificial-intelligence-against-modern-slavery/#:~:text=Project%20AIMS%20(Artificial%20Intelligence%20against%20Modern%20Slavery)%20%2D%20The%20Future%20Society&text=The%20AI%20Initiative%20is%20dedicated,and%20long%2Dterm%20governance%20challenges.

[3] https://thefuturesociety.org/2020/07/28/caiac-alliance-launch/

[4] https://thefuturesociety.org/2019/11/14/the-ai-commons/

[5] https://thefuturesociety.org/2019/11/15/global-data-access-framework-gdaf/

[6] https://thefuturesociety.org/2019/02/10/global-governance-of-ai-forum-ggaf/

[7] https://thefuturesociety.org/2020/08/31/development-of-rwandas-national-artificial-intelligence-policy/

[8] https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/ai/ey-bridging-ais-trust-gaps-report-2020.pdf

[9] https://thefuturesociety.org/2020/12/17/report-release-with-the-global-partnership-on-ai/

[10] https://www.oecd.ai/dashboards/countries/Japan

[11] OECD.AI Policy Observatory, National Strategies & Policies, Japan, https://www.oecd.ai/dashboards/countries/Japan. Accessed on December 22, 2020.

[12] https://www.forbes.com/sites/markminevich/2020/03/16/women-are-the-key-to-scaling-up-ai-and-data-science/?sh=16668a935ac8

[13] https://www.designnews.com/electronics-test/theres-diversity-crisis-ai-industry

[14] https://www.fastcompany.com/90558017/half-of-women-who-work-in-tech-do-this-surprising-thing-by-age-35#:~:text=Fifty%20percent%20of%20women%20in,and%20Girls%20Who%20Code%20reveals.

[15] https://thenextweb.com/neural/2020/10/27/institutional-bias-and-lower-acceptance-rates-for-women-inside-the-ai-conference-review-process-syndication/

[16] https://www.honeypot.io/women-in-tech-2018

[17] https://www.womens-forum.com/wp-content/uploads/2020/04/WomenAI_DaringCircle_GMPressRelease_FINAL.pdf

[18] https://www.womeninai.co/

[19] https://100brilliantwomeninaiethics.com/

[20] https://ai-4-all.org/

[21] https://www.elementsofai.com/

[22] https://www.europeanwomenintech.com/

[23] https://www.womenofsiliconvalley.com/

[24] https://www.women-in-technology.com/

[25] https://www.womeninai.co/

Yolanda Lannquist

Yolanda Lannquist is Head of Research & Advisory at The Future Society, a nonprofit 'think-and-do-tank' with the mission to advance the responsible adoption of artificial intelligence (AI) and emerging technologies for the benefit of humanity. Yolanda leads projects in AI governance and policy including developing national AI strategies, harnessing AI for Sustainable Development Goals, and mitigating the ethical, safety and societal impacts of AI. She directs AI policy courses at Sciences Po Paris School of International Affairs and IE University in Madrid. Previously, Yolanda advised Fortune 500 multinationals on machine learning, innovation and market entry strategy as a business strategy consultant. She also worked on digital trade and regulation at the U.S. Embassy in Paris and authored several reports on global labor markets and human capital at The Conference Board in New York. Yolanda has a Master in Public Policy from Harvard University's Kennedy School of Government and Bachelors in Economics and European Studies from Columbia University in New York.

Adriana Bora

Adriana Bora is an AI Policy Researcher and Project Manager at The Future Society, a member of the MIT Computational Law Report Task Force on Modern Slavery, and a contributor to The Good AI. Through her research, Adriana studies how augmented intelligence can accelerate the eradication of modern slavery. She is applying machine learning in analyzing and benchmarking the businesses' reports published following the Modern Slavery Act from the UK and Australia. Adriana holds a Masters Degree in International Public Management at Sciences Po Paris School of International Affairs. She has also studied for a year at the University of Hong Kong and holds a diploma in International Relations and Advanced Quantitative Methods from the University of Essex in the UK.

Niki Iliadis

Niki is an international specialist in AI governance & policy. As Senior AI Policy Researcher and Project Manager at The Future Society, Niki supports the development and implementation of projects ensuring the development and deployment of AI benefits all of humanity. She recently led a landscape review and analysis of future opportunities in the Responsible AI ecosystem for the Global Partnership on AI. Her interests include agile governance approaches, AI's impact on SDGs, civic participation in AI decision-making, and AI's impact on children. Prior to joining TFS, she led the delivery of the All-Party Parliamentary Group on Artificial Intelligence (APPG AI), an initiative in the UK Parliament gathering evidence on the economic and socio-ethical impact of AI, and using that to shape policy. Previously, Niki's experiences include working for a global foundation to bridge the policy implementation gap (Centre for Public Impact – a BCG Foundation), a primary research firm to gather due diligence across industries (Third Bridge), an international think tank to develop projects in youth engagement (Strategy International), an education-focused NGO (Centre for Democracy and Reconciliation in Southeast Europe), and the Embassy of the United States. She holds a BSc in Political Science from UC Berkeley and an MSc in Public Management from London School of Economics.

The Diversity and Ethics Crisis in AI

Mia Shah-Dand Women in AI Ethics™

Introduction

Diversity and ethics in Artificial Intelligence (AI) are often treated as two separate and unrelated issues. However, lack of diversity in AI and the wider tech industry are increasingly showing up in AI and Machine Learning (ML) models and systems. The term 'AI ethics' is used interchangeably with 'ethical AI' or 'responsible AI', all of which describe the frameworks to address current and potential unintended consequences of AI/ML systems. Governance and oversight of these technologies is essential to reduce harm, increase accountability, and collectively shape the types of outcomes we want to see from AI.

Diversity takes many different forms. Groups that do not conform to the mainstream norms or belong to dominant power groups are often relegated to a powerless position and their role is effectively marginalized. Our definition of diversity considers the intersectionality of different dimensions, including but not limited to gender, race, orientation, ability, age, region, religion, social-economic, and others as well as relative power structure in our work.

The Diversity Crisis in AI

The tech industry is notorious for its lack of diversity and despite years of talk, the figures are still dismal. Less than 15% of AI researchers at big tech are women [1] and progress is even worse on racial representation as less than 5% of the workforce at tech giants is Black [2]. We see this inequity repeated at AI conferences where only 18% of authors are women [3] and also in academia where 80% of AI professors at leading universities are men [4]. Recent studies [5] show that AI is also predominantly portrayed as white in media, which includes "humanoid robots, chatbots and virtual assistants, stock images of AI, and portrayals of AI in film and television."

Lack of diversity extends beyond race and gender. Despite many technological advancements over the past decade, folks with disabilities are still excluded from participation as only 10% of disabled people have access to so-called assistive technologies [6].

Datasets used to train machine learning models are especially problematic as research shows racism and sexism are embedded in historical AI datasets used for training machine learning models. MIT had to remove a huge dataset [7] that taught AI systems to use racist and misogynistic slurs. In 2009, Researchers at Princeton and Stanford created a database called ImageNet [8] by collecting photos from websites like Flickr and used low wage workers at Amazon Mechanical Turk to categorize the photos. This online image database had to remove 600,000 pictures [9] after an art project revealed the system's racist bias.

More than 50% of images in popular datasets come from the U.S. and U.K [10]. The data gathering process for machine learning models is ethically questionable [11] as there is very little transparency around who is using our dataEven the elite institutions who have positioned themselves as AI ethics gatekeepers are themselves responsible for many ethical lapses [12].

Duke University terminated a dataset of surveillance [13] footage used for research and development of video tracking systems and low-resolution facial recognition, after public backlash [14] over privacy violations. Google's AI company, DeepMind came under fire [15] for its deal with UK's health service to transfer 1.6 million patient records which [16] was deemed to have "inappropriate legal basis" according to a data privacy watchdog. Even after the datasets are taken down, they still continue to live [17] on in many AI models

and systems that were built using the flawed datasets.

Lack of diversity is reflected in academia and especially in AI research, which is concentrated in the hands of the privileged and powerful techno-elites. At Ivy Leagues, more students are from top 1% of America's richest families than from bottom 60% [18] which further exacerbates inequities in the tech industry.

Majority of papers at prestigious AI conferences like NeurIPS 2020 [19] and ICML 2020 [20] are from Google, Stanford, and MIT. U.S. participated in more than 4 times (1186) the number of papers submitted by China (259) and the U.K. (205) while Japan was at #12 with 38 papers at NeurIPS [21]. Similarly, U.S. accounted for 728 accepted papers, which was 75% of all the papers submitted at ICML 2020 [22] while Japan was in the top 10 with 31 papers.

Lack of Diversity Leads to Unethical Outcomes

Lack of diversity at tech companies combined with racism and sexism embedded in historical datasets is further amplified in the output from these AI models, which are widely used across a broad range of industries from retail, healthcare to law enforcement. While accuracy of machine learning models is not the same as ethics the two are often conflated and it's fair to say that models trained on biased datasets and algorithms are also likely to be less accurate for marginalized groups.

In October 2019, a study [23] found that algorithms grossly underestimated the number of Black patients who needed extra care because it used health costs as a proxy for health needs. Historically, there was less money spent on Black patients than equally sick white patients, so the algorithms mirrored and amplified the racist bias embedded in its training datasets. Another study [24] also found that an algorithm used for assessing a person's kidney function "often overestimated the health of Black patients, resulting in them receiving less specialized care, or worse, keeping them from getting placed on a kidney transplant waitlist."

This racist bias is reflected in experience of and subsequent research by Dr.Joy Buolamwini [25] who tested multiple facial analysis demos and found that two didn't detect her face and others misgendered her. According to a BBC investigation last year [26], the U.K. passport photo checker on the application website also showed bias against dark-skinned women.

Gender bias shows up in discriminatory financial services algorithms, as highlighted by a recent incident with Apple card where female applicant was given a lower credit limit [27] compared to her male partner with a similar credit score.

Researchers found [28] that machine-learning software trained on prominent image datasets both mirrored and amplified the historical biases in the datasets by giving more weight to stereotypical associations. This sexist and racist bias shows up in search algorithms [29] and in AI assistant voices [30], which reinforce "obliging, docile and eager-to-please helpers" stereotype. Gender bias in the tech industry data used for training recruiting machine learning models [31] results in discrimination against women candidates.

Last month, news about Dr. Timnit Gebru, the eminent and beloved AI Ethics scholar and researcher fired [32] by Google, for co-authoring a research paper exploring the potential risk of large language models [33], roiled the industry and highlighted the unsavory reality of being Black and ethical in a space dominated by powerful white men. It revived traumatic memories for many women of color who have faced gaslighting, exploitation, and erasure in the toxic tech industry. It also brought to light the broader issue of credibility and objectivity of AI ethics research funded by big tech [34].

A recent study [35] unearthed that a significant number of faculty at top universities have received some form of financial support from big tech. Insidious influence of big tech shows up in framing of AI Ethics research, most of which is focused on solving ethical issues in such a way that AI development can continue unabated. Much of it is centered around risk mitigation on behalf of and influenced by tech companies rather than well-being of marginalized communities.

In March 2019, Stanford launched the Institute for Human-Centered AI [36] with an advisory council glittering with Silicon Valley's brightest names, a noble objective of "to learn, build, invent and scale with purpose, intention and a

human-centered approach," and an ambitious fundraising goal of over $1 billion.

This new institute kicked off with glowing media and industry reviews, when someone noticed a glaring omission. Chad Loder [37] pointed out that the 121 faculty members listed were overwhelmingly white and male, and not one was Black [38].

Rather than acknowledging the existence of algorithmic racism as a consequence of anti-Blackness at the elite universities that receive much of the funding and investment for computer science education and innovation, or the racism at tech companies that focus their college recruitment at these schools, we act as though these technological outcomes are somehow separate from the environments in which technology is built.

Every so often big tech companies and elite institutions trot out AI Ethics luminaries to make an eloquent speech on the need for more ethical AI but when experts like Dr. Gebru point out the ethical flaws of these technologies [39], they are attacked, discredited, and discarded with impunity. The inevitable conclusion is that AI Ethics initiatives by big tech are designed to make problematic tech more palatable and they are used merely as a smokescreen to hide their transgressions.

Whether it is sentencing recommendation or crime prediction, models based on flawed algorithms and biased datasets reinforce prejudices of the past and have a significant impact on people's lives. Given the pervasiveness of bias in Artificial Intelligence (AI) algorithms and the existential threat it poses to marginalized groups there is an urgent need for an open discussion and concrete action to address the perils of unchecked AI.

Women in AI Ethics™

The Women in AI Ethics™ initiative was launched in 2018 with publication of the first 100 Brilliant Women in AI Ethics™ [40] list created by Mia Shah-Dand, CEO and Founder of Lighthouse3 [41] an emerging technology research and advisory firm based in Oakland, California. Her intent was to increase recognition, representation, and empowerment of talented women in this space.

This list [42] is now published annually and supplemented with an online directory [43] to highlight rising stars as well as aspiring women from non-technical fields as part of an ongoing effort to make AI more diverse and accessible for everyone. In July 2020, Women in AI Ethics™ reaffirmed their commitment to diversity & inclusion in AI by assuming a vocal advocacy role as staying neutral in times of social and racial injustice is ethically and morally unacceptable. Towards that end, this initiative now centers the experiences of marginalized and underrepresented groups in all their work.

In November 2020, Women in AI Ethics™ [44] (WAIE) became a fiscally sponsored project of Social Good Fund, a California nonprofit corporation and registered 501(c)(3) organization with a mission to increase recognition, representation, and empowerment of brilliant women in this space who are working hard to save humanity from the dark side of AI. It is funded by Lighthouse3, donations, and ticket sales. Donations are tax deductible to the extent allowed by law and used to fund mission-aligned activities, which includes but is not limited to providing free AI Ethics career resources, hosting community events, and funding other initiatives to support women in this space.

On December 3-5, 2020, 100 Brilliant Women in AI Ethics™ [45] brought together over 25 activists, social scientists, data scientists, engineers, artists, researchers, policy makers, anthropologists, and other brilliant women from technical and non-technical backgrounds from around the world. This unique event showcased how diversity leads to more ethical and inclusive AI that works for all of humanity. This event included voices from Indigenous communities and non-English speaking regions to reflect non-western perspectives.

Impact of COVID-19

"The unfolding impacts of COVID-19 reveal just how many communities of women, and the families that depend on their earnings, are bearing the brunt of the longstanding gaps and underinvestment in our workplace laws, economic and social infrastructure, and policy choices that failed to center the needs of women, people of color, and families with low and moderate incomes." ~National Women's Law Center

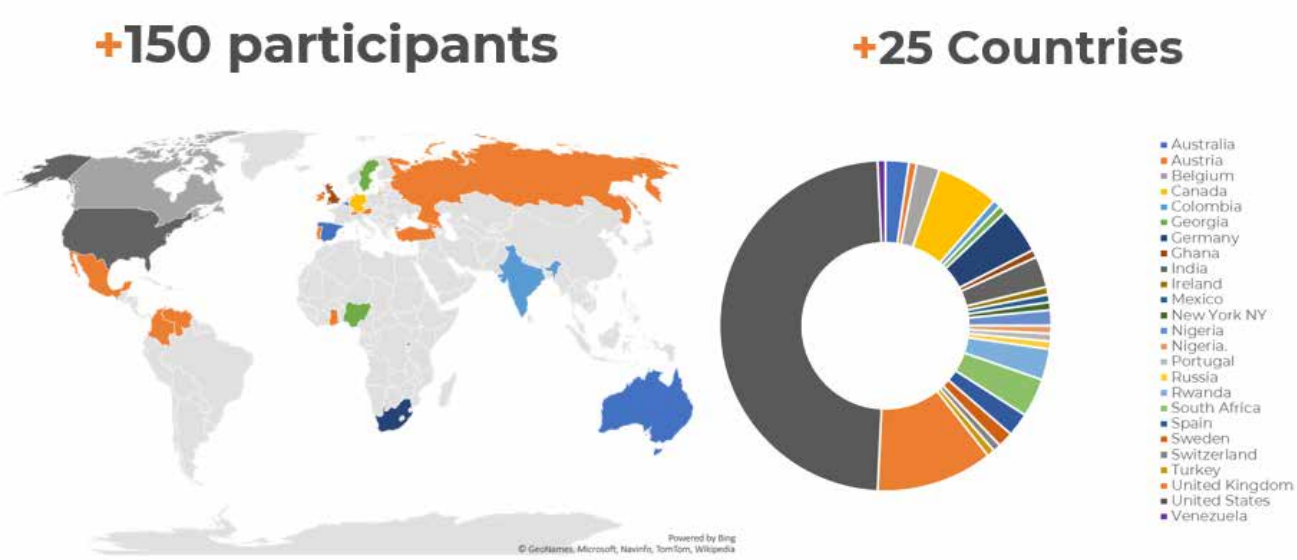
According to the ILO report 2018 [46], women do the bulk of unpaid care work in homes across the world, a workload that has intensified during this pandemic. This includes older women caring for frail partners and grandchildren. For those women able to work from home, the sudden need during this pandemic to home-school children has created a double (or triple) shift.

In a study by the The National Center for Transgender Equality [47], transgender and non-binary people had double the rate of unemployment than that of the general population. This pandemic will further reduce their access to paid opportunities.

One in four women are considering a career step-back [48] and female researchers are submitting fewer journal [49] articles than their male peers because they are overwhelmed and distracted by challenges of dealing with COVID-19. These backward moves will have negative impact on their career trajectory and one from which they may never recover.
Fewer opportunities for women and minorities translates into lower ethical accountability for companies as power shifts to the latter. It's also during times of crisis that humanity is most at risk from unethical use of powerful technologies like Artificial Intelligence (AI).

AI Ethics Mentoring Program

To support women and non-binary folks in AI Ethics and empower them with more resources during this critical time, Women in AI Ethics™ and Lighthouse3 has launched a global AI Ethics mentoring program [50] in the summer of 2020 for a limited time to provide support and guidance for vulnerable communities, which included women and non-binary people. Within 6months, this program had over 150 participants from 25 countries.



Source: https://lighthouse3.com/mentoring/

The feedback from the AI Ethics mentoring program (which is currently on hold) has been overwhelmingly positive and here are some key highlights from our participants.

- Women feel unsupported in the male-dominated AI/tech space.
- Many lost their job, internship, or academic opportunity because of COVID-19.
- Some felt their university or employer didn't provide adequate mentoring support.
- Students appreciated advice from experienced mentors on their research projects.
- It was helpful for those in academia to learn about industry perspective on AI/Machine Learning.
- Non-technical mentees appreciated advice from non-technical mentors on their career journey.
- Mentees were also motivated to help others as a mentor, which created a virtuous helpful cycle.

Here are some quotes from the mentors and mentees in the AI Ethics mentoring program:

> "Mentoring is a two-way benefit and an inspiring experience."
> "It means a lot to feel supported in this space, where we are typically underrepresented."
> "The unique expertise of my mentor both on an ethical AI and career development level contributed to a holistic picture on complex issues."

AI Ethics Framework

After an extensive evaluation of the AI ecosystem and broad implications of AI, Women in AI Ethics™ uses a comprehensive framework for AI + Ethics, developed by Lighthouse3 with 6 key focus areas, which consider all ethical implications of AI beyond just technology development. It is intentionally designed to expand critical discussions on the ethics of AI to include diverse perspectives from non-technical disciplines. We firmly believe that we cannot have a meaningful discussion about the ethics of AI without including marginalized and underrepresented groups in these critical conversations. It is intentionally designed to expand critical discussions on the ethics of AI to include diverse perspectives from non-technical disciplines.

COMMERCIAL SYSTEMS + SAFETY
Provide guardrails and guidance for development of commercial autonomous systems, artificial agents that are ethical and safe.

DIGITAL + PHYSICAL SAFETY
Establish safeguards against threats to digital and physical safety through deep-fakes, bots, autonomous weapons, and bio/neural AI.

FAIRNESS + ACCOUNTABILITY
Ensure AI/ML models are transparent, explainable, fair, and accountable so they do not reflect and amplify the biases of their creators.

PRIVACY AND DATA RIGHTS
Ensure that data for AI/ML is gathered (surveillance), shared (consent) ethically and secured against unethical uses (security).

ROLES + RIGHTS
Mitigate disruption of human lives, rights, roles, and relationships as AI replaces humans through automation & human-like systems.

SOCIETY + SUSTAINABILITY
Promote use of AI for social good, through access for marginalized groups, mitigation of environmental impact, and development of public interest policies.

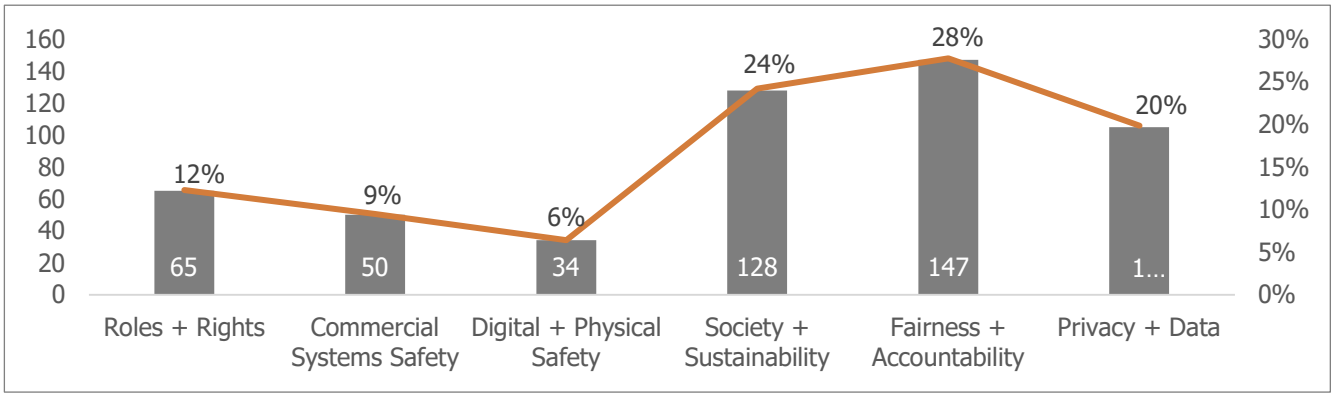
These dimensions are **not** mutually exclusive but rather this framework allows us to address each significant issue individually without losing sight of the interconnectedness of the different parts. For example: AI automation causes jobs displacement so while it's important to look at ethics of intelligent systems, we must not lose sight of the displaced workers who will need skills and training to navigate this new world.

Women in AI Ethics ™ – Areas of Focus & Specialization

Based on review of 500+ profiles in the Women in AI Ethics™ directory (self-reported or through their professional website) we found that majority of women in this space are focused on 3 key areas of AI ethics: Society + Sustainability, (Algorithmic) Fairness + Accountability, and Privacy + Data.

Given significant funding of AI research by big tech companies, overwhelming media coverage of algorithmic bias, and corporate interest in risk mitigation could partly explain why there is more focus on Fairness + Accountability to ensure that the algorithmic models are bias-free and marketable.

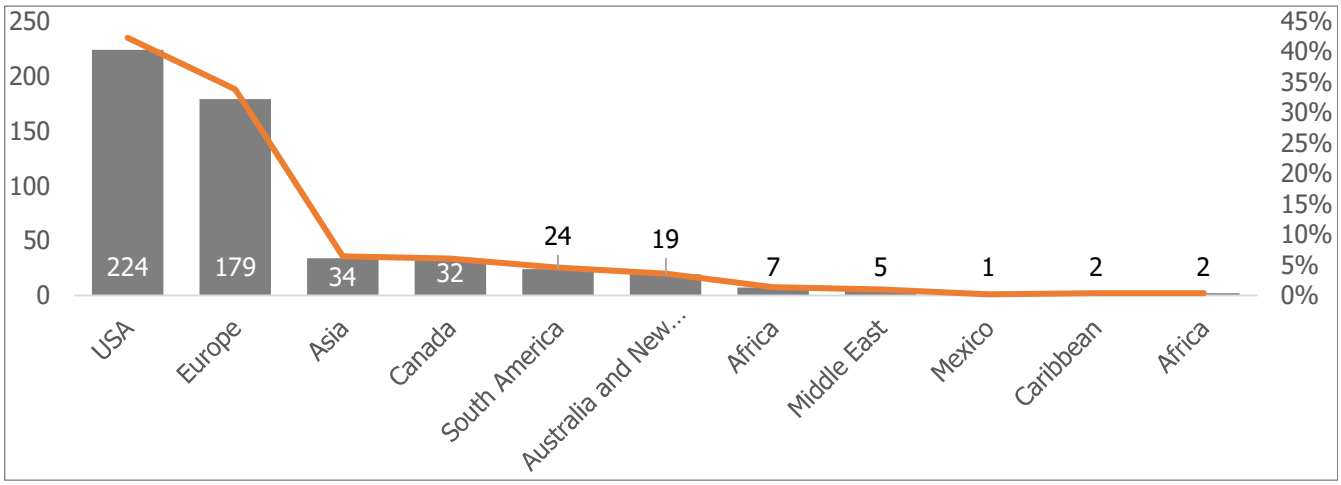
Second-most popular category is Society + Sustainability, which is gaining in popularity as it's a good first step for non-technical women to step into the world of AI Ethics and naturally aligns with other social good efforts to improve access for marginalized groups, mitigation of environmental impact, and development of public interest policies. Further analysis is needed to provide more actionable insights.



Source: https://womeninaiethics.org/ Note: Some may have multiple areas of focus.

Women in AI Ethics™ – Location

We found overwhelming representation of women who indicated they were based in the U.S. with over 40% of the total number of profiles in the WAIE directory, followed by Europe (including U.K.) at 34% while representation from Asia and other countries is still in single digits. This could be attributed to lack of awareness or lack of women in this space. Regardless of the reason, it's clear we need to do more to increase participation of women from Japan and other parts of Asia in discussions on the ethics and governance of AI.



Source: https://womeninaiethics.org/ Note: Some maybe in multiple locations

Pathway to Ethical and Inclusive Future.

Technologies reflect the priorities and ethics of those building and funding them. We must start by acknowledging that technological outcomes are not separate from the environments in which these technologies are built. To build more ethical AI technologies, governments must act as role models by nominating women to key decision-making roles on these critical issues. It's crucial to include representation from marginalized communities impacted by output of AI technologies in key discussions and decisions to avoid harmful blind spots.

The road to ethical AI and diverse AI starts with being intentional about diversity and ethics followed up with the right decisions to support that goal. Being purposeful about not harming marginalized communities means organizations and governments need to center their experiences in the development process. They should have the moral fortitude to ban questionable uses of technologies that harm marginalized communities even if said technologies were developed using ethical design principles.

An educated and well-informed citizenry is the best defense against misinformation about AI. There is an urgent need to

shift our focus away from developers and purveyors of these technologies and towards making more investments in educating everyone touched by these technologies. Finland launched a free online course [51] to educate all its citizens on the basics of AI and is a great model to follow for nations in their quest for more inclusive AI. Investment in translation of AI curriculum from English to Japanese and vice versa will also make it more accessible, bring down barriers, and encourage further collaboration.

Research [52] has shown a linear relationship between racial, ethnic diversity and better financial performance. Companies with more diverse workforces perform better financially. Lowering the barriers to AI education will ensure that women and marginalized communities have access to resources and opportunities that were previously closed off to them. But it's not enough to open the door to access but there should be adequate protections for women and other marginalized communities to avoid the revolving door of talent because of toxic environments. Both private and public sector organizations should be encouraged to build welcoming environments for women in traditionally male-dominated roles.

Creation of meaningful legislation and policies to eliminate or at the very least, minimize harms to marginalized communities is essential as is the inclusion of marginalized voices in development of AI technologies. Rather than just exploiting consumers for labor and data, big tech should be nudged to share benefits from development of these technologies with everyone instead of restricting access to the wealthy and privileged. Diversity can help identify the blindspots that may not visible otherwise to the privileged.

Establish protocols for tracking and measuring diversity at tech companies and conferences. Encourage and reward women working in this space. Nominate Japanese women for the WAIE directory and encourage them to apply for speaking roles at prestigious international conferences. Partner with initiatives like WAIE to provide support and mentoring for Japanese women who may otherwise feel isolated.

Redefine who can be an "AI expert" and include women from non-academic and non-technical backgrounds in AI discussions and events. Make it easier for women to go back to school for further education and training in AI.

Diversity is the gateway to a more ethical and inclusive future for all of humanity.

[1] https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/

[2] https://ainowinstitute.org/discriminatingsystems.html

[3] https://jfgagne.ai/talent-2019/

[4] 80% of AI professors at leading universities are men https://hai.stanford.edu/sites/default/files/2020-10/AI_Index_2018_Annual_Report.pdf

[5] https://philpapers.org/rec/CAVTWO

[6] https://ainowinstitute.org/discriminatingsystems.html

[7] https://www.theregister.com/2020/07/01/mit_dataset_removed/

[8] http://image-net.org/

[9] https://www.artsy.net/news/artsy-editorial-online-image-database-will-remove-600-000-pictures-art-project-revealed-systems-racist-bias

[10] https://arxiv.org/pdf/1711.08536.pdf

[11] https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e

[12] https://gritdaily.com/where-is-the-accountability-for-ai-ethics-gatekeepers/

[13] https://www.dukechronicle.com/article/2019/06/duke-university-facial-recognition-data-set-study-surveillance-video-students-china-uyghur

[14] https://www.dukechronicle.com/article/2019/06/duke-university-facial-recognition-data-set-study-surveillance-video-students-china-uyghur

[15] https://www.theguardian.com/technology/2017/may/16/google-deepmind-16m-patient-record-deal-inappropriate-data-guardian-royal-free

[16] https://www.theguardian.com/technology/2016/may/04/google-deepmind-access-healthcare-data-patients

[17] https://freedom-to-tinker.com/2020/10/21/facial-recognition-datasets-are-being-widely-used-despite-being-taken-down-due-to-ethical-concerns-heres-how/

[18] https://www.nytimes.com/interactive/2017/01/18/upshot/some-colleges-have-more-students-from-the-top-1-percent-than-the-bottom-60.html

[19] https://neurips.cc/

[20] https://icml.cc/

[21] https://medium.com/criteo-engineering/neurips-2020-comprehensive-analysis-of-authors-organizations-and-countries-a1b55a08132e

[22] https://medium.com/criteo-engineering/icml-2020-comprehensive-analysis-of-authors-organizations-and-countries-c4d1bb847fde

[23] https://science.sciencemag.org/content/366/6464/447

[24] https://www.mic.com/p/how-a-racist-algorithm-kept-black-patients-from-getting-kidney-transplants-40315478

[25] https://www.media.mit.edu/projects/gender-shades/overview/

[26] https://www.bbc.com/news/technology-54349538

[27] https://hbswk.hbs.edu/item/gender-bias-complaints-against-apple-card-signal-a-dark-side-to-fintech

[28] https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/

[29] https://www.salon.com/2013/10/20/new_campaign_uses_real_google_searches_to_expose_how_the_world_talks_about_women/

[30] https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

[31] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08Gus-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[32] https://miad.medium.com/dr-timnit-gebru-big-tech-and-the-ai-ethics-smokescreen-45eb03d1fe6d

[33] https://www.bloomberg.com/news/newsletters/2020-12-08/how-timnit-gebru-s-academic-paper-set-off-a-firestorm-at-google

[34] https://upfromthecracks.medium.com/on-the-moral-collapse-of-ai-ethics-791cbc7df872

[35] https://www.wired.com/story/top-ai-researchers-financial-backing-big-tech/

[36] https://hai.stanford.edu/education

[37] https://twitter.com/chadloder/status/1108588849503109120

[38] https://gizmodo.com/stanfords-new-institute-to-ensure-ai-is-representative-1833464337

[39] https://www.theverge.com/2020/12/5/22155985/paper-timnit-gebru-fired-google-large-language-models-search-ai

[40] https://lighthouse3.com/our-blog/100-brilliant-women-in-ai-ethics-you-should-follow-in-2019-and-beyond/

[41] https://lighthouse3.com/

[42] https://100brilliantwomeninaiethics.com/the-list/

[43] https://womeninaiethics.org/directory.php

[44] https://womeninaiethics.org

[45] https://100brilliantwomeninaiethics.com/

[46] https://data.unwomen.org/features/covid-19-and-gender-what-do-we-know-what-do-we-need-know

[47] https://transequality.org/issues/resources/national-transgender-discrimination-survey-full-report

[48] https://qz.com/work/1911086/covid-19-has-working-mothers-considering-a-career-step-back/

[49] https://hssonline.org/isis-submissions-and-gender/

[50] https://womeninaiethics.org/mentoring/

[51] https://course.elementsofai.com/

[52] https://www.mckinsey.com/business-functions/organization/our-insights/why-diversity-matters

Mia Shah-Dand

Mia Shah-Dand is the CEO of Lighthouse3, a research and advisory firm based in Oakland, California. She is the founder of Women in AI Ethics™ initiative dedicated to increasing recognition, representation, and empowerment of talented women in this space. Mia is the creator of 100 Brilliant Women in AI Ethics™ annual list and Women in AI Ethics™ online directory, a resource to increase representation of women at AI conferences and in the tech industry. Mia is on the Board of Directors for the United Nations Association - USA San Francisco Chapter and Advisory Board for Carnegie Council's AI & Equality Initiative.

The greatest scientific challenge

Olga Afanasjeva GoodAI

A bold vision

At GoodAI our mission is to build safe general artificial intelligence - as fast as possible - to help humanity and understand the universe.

This pursuit is not just about the technology, which is simply a means to an end. It's about creating a future society where humans will be capable of overcoming the insurmountable challenges we face today in medicine, engineering, and a range of other fields. At GoodAI we recognize that this pursuit is bigger than one company. In this article, I'd like to offer our perspective on how to encourage a community effort. Romantically, one could compare it to a vision of a renaissance movement - influential, lasting, interdisciplinary, and seeing institutional and capital support.

By general AI, or AGI, we mean a kind of AI that would augment the intelligence of humans in a broad sense. Unlike present-day AI technology which is good at solving specific tasks it was designed for, a general AI would be able to solve any new problems in creative ways, and independently set and then tackle challenging research and engineering goals. Another way to define general AI is, an AI that would be able to create specialized solutions for different problems without the need for tedious human effort.

We believe that achieving a safe, beneficial general AI poses the most important challenge for humanity. Once achieved, AGI would provide unprecedented leverage across fields, since creative intelligence won't be scarce anymore, and would be much easier to scale (instead of having to wait for generations of engineers and scientists to mature, we could put a few humans and a few computers together).

Such a grand vision asks for special focus on two questions:

1) how can we scale up the research to get to the goal faster, without compromising safety?
2) how can we set the right governance mechanisms, including rigorous testing, to make sure different parties can efficiently cooperate together towards a technology that benefits humanity as a whole?

Both require thinking beyond traditional paradigms of how to do research, and developing much more robust models of cooperation. A premise here is that AGI would be an enabling technology much more potent than anything we've seen before. That leads to risks of a "winner takes all" dynamic of fierce competition, shortcuts on safety for the sake of speed, growing inequality, and monopoly on access to AGI technology.

Expanding the traditional paradigms of how to cooperate in research and development

A roadmap to general AI

In a project like AGI, big picture thinking is particularly important. Nobody has made it to AGI before, the reference points are few and are rather long-shot analogies (for example, the human brain or biological evolution). To move forward without losing perspective and getting stuck at the wrong problems, you would want to have a roadmap outlined, and maintain the right balance between reasonable doubt and commitment to a certain roadmap. The commitment part can be very challenging when it comes to uncharted territory problems, where you can easily get discouraged by preliminary results. Allocating resources is also a challenge when a lot of alternative paths quite often look equally worth pursuing.

We defined our research roadmap by asking, what are the main features of intelligence that we would want our agents to have, and what are the possible pathways to get to these features. We looked for inspiration into how humans and other intelligent organisms can gradually acquire new skills and abilities with elegant efficiency. They build upon and creatively reuse existing knowledge to solve novel tasks, without forgetting useful heuristics or re-writing the entire brain (most deep learning systems need to be retrained for separate tasks which isn't very efficient).

This thinking led us to the development of Badger architecture, a unifying framework for our research, where the key principle is modular life-long learning [1].

Collaborating along the road

We see other researchers working in a number of promising directions, a lot of these directions are related and complementary. Together with the research and engineering we do in our company, our goal is to create conditions for groups and individuals to work in synergy towards a joint mission.

In 2020, we announced GoodAI Grants. Through this program, we provide funding and suggest topics for applicants to tackle. These topics are based on our current state of research in the GoodAI team, and many of them mirror milestones our team is working on. At the same time, we stay open to proposals outside the direction of our Badger research framework. Anyone, an individual or a group irrespective of their background, is eligible to submit a grant proposal.

The GoodAI Grants program aims to create comfortable conditions for cooperation and knowledge-sharing. Participation in the community should be easy to balance with other obligations in academia or the private sector.

An important goal of the program is to focus the effort and improve the roadmap to more human-like AI through collaboration: new findings adjust the roadmap, which in turn generates new research questions. We also felt there was a need for an attractor for bold ideas that would incentivize people to work on really hard problems and reward efforts that have long-term benefits.

So far, we have awarded over $300k through the Grants (all recipients are soon to be announced). The program runs continuously and is open to applications at any time.

Another pillar of our collaborative endeavor is the citizen science project, GoodAI's **General AI Challenge.** Our aim has been to open the door for anyone, irrespective of field or background, to participate in what we believe is an interdisciplinary discourse. It's also a meta mechanism: through the General AI Challenge, we can both do research, as well as discover better ways to do research.

Together with posing specific technical questions, the mission of the Challenge is to search for the best possible structures and mechanisms for AI governance, cooperation, and trust-building. This is where people with a background in social sciences and humanities are most encouraged to join the conversation (which might still be perceived as a techy-niche), proving that the AGI development is a truly universal, impact-focused effort.

One of the rounds of the Challenge focused on the question of an AI Race where:
Key stakeholders, including the developers, may ignore or underestimate safety procedures, or agreements, in favor of faster utilization,
The fruits of the technology won't be shared by the majority of people to benefit humanity, but only by a selected few.

The primary objective was to find a solution or set of solutions for mitigating the risks associated with the AI race. The secondary objectives were to create discussion around the topic in order to gain a better understanding of the nature of the AI race, raise awareness of the race, and to get as diverse an idea pool as possible. This round received 194 registrations from 41 different countries and a total of 59 submissions [source: General AI Challenge Website]. You can check out the awarded submissions here [2].

One of the recurring themes in submissions was creating meaningful cooperations through transparency and trust-building. For the next challenge, we would like to narrow down the question and ask, what are the optimal mechanisms for fair sharing of intellectual property among diverse

cooperators? What is a cooperation model that would protect the proprietary rights and interests of parties, and at the same time provide a fair title to joint results?

In the present day, this is quite often solved through two extremes, open-source or complicated tailor-made contracts which can in themselves present a hurdle towards cooperation. Power balance might come into play when one party could be at a disadvantage simply because they have less legal backing. The ideal outcome is where none of the parties are limited, but are equally protected and presented with equal opportunities to utilize the results, not based on power dynamics (e.g. who is a better negotiator of contracts).

The design of this round of the General AI Challenge is still in the making. As submissions, we imagine example legal frameworks, case studies, or perhaps even designs of plug and play platforms for generating meaningful contracts. If you have ideas or would like to be involved in the design phase of this round, please don't hesitate to get in touch.

My message to the reader of this article is an encouragement to think creatively about how to achieve bold goals, and I hope that if you're interested in developing human-like AIs, you would find our approaches and our programs interesting and engaging. Get in touch if you have suggestions for us, or if you would like to become part of the collaborative effort.

---

[1] The modular aspect is expressed in the architecture through a network of identical agents. The life-long learning means that the network will be capable of adapting to a growing (open-ended) range of new and unseen tasks while being able to reuse knowledge acquired in previous tasks. The algorithm that is run by individual Badger experts will be discovered through meta-learning.

We expect the design principles of Badger architecture to be its key advantages. The modular approach should enable scaling beyond what is possible for a monolithic system and the focus on life-long learning will allow for incremental, piece-wise learning, driving down the demand for training data. [Source: https://www.goodai.com/badger-architecture/]

[2] https://www.general-ai-challenge.org/active-rounds

Olga Afanasjeva

Olga is an AI evangelist with a background in arts and social sciences, pursuing her passion for discovery and redefining the limits of what is possible. She leads the day to day activities of international research and development organization GoodAI. GoodAI was founded with a $10M investment by CEO/CTO Marek Rosa in January 2014 with the goal to develop general artificial intelligence - as fast as possible - to help humanity and understand the universe.

Commentary: From AI Principles to Practice: Lessons for Japan to Learn from International Activities

Arisa Ema Institute for Future Initiatives, The University of Tokyo / RIKEN

1. Introduction

"From Principle to Practice" is the title of the first chapter of Ethically Aligned Design, 1st Edition, published by the Institute of Electrical and Electronics Engineers (IEEE) Global Initiative on the Ethics of A/IS in March 2019. After three to four years of international multi-stakeholder discussions, various issues related to the use of Artificial Intelligence (AI), which started around 2015, have led to the realization of a shared set of general principles.

In spring 2019, important documents were also released in Japan and Europe: the "Social Principles of Human-Centric AI" was released in March by the Japanese Cabinet Secretariat, and the "Ethics Guidelines for Trustworthy Artificial Intelligence" was published in April by the High-Level Expert Group on Artificial Intelligence (AI HLEG), a group of independent experts created by the European Commission. Countless other principles are currently published around the world and, as reported in the article by Ms. Lannquist and her colleagues, the OECD operates the AI Observatory website as a centralized portal. The website records more than 300 AI policies and initiatives in more than 60 countries and regions.

Numerous articles have been published on the values shared by these principles. For example, a report from Harvard University in the United States [1] [2] and a review prepared by researchers from the Chinese Academy of Sciences, shared almost the same values (see Table 1).

[Talbe 1] :Important values in the principles of artificial intelligence. The common values between the Chinese Academy of Sciences and the Berkman Klein Center at Harvard University are reported with the same color.

Chinese Academy of Sciences(2018)	Berkman Klein Center at Harvard University(2019)
Humanity	International Human Rights
Collaboration	Promotion of Human Values
Share	Professional Responsibility
Fairness	Human Control of Technology
Transparency	Fairness and Non-discrimination
Privacy	Transparency and Explainability
Security	Safety and Security
Safety	Accountability
Accountability	Privacy
AGI	—

The values listed in Table 1, such as safety, security, and privacy, represent issues traditionally related to information technology. On the other hand, fairness, accountability, and transparency are three values emphasized in the AI debate, and are indicated by the acronym FAT. This is sometimes referred to as FATE, when explainability is also considered.

Unfortunately there are some cases where the AI output is

unfair and discriminatory due to biases in data training and algorithms, and designers may be (unconsciously) influenced by the discrimination and prejudices of society. Although these problems mainly occur in foreign countries, as Japan promotes the use of AI, people could be subjected to discriminations and prejudices in our country as well.

These principles are not irrevocable. How to use these in practice has been clearly defined since 2019. In response to this, the academia, research institutes, and global companies are introducing frameworks to verify and guarantee security and fairness in the development process of AI systems and services [3].

Some national policies are encouraging this trend. For example, some countries have started to include AI governance rules for government procurement. Some countries and regions are also proposing the use of frameworks, toolkits, and certifications for safety and fairness self-monitoring in some organizations. It is undeniable that researchers and companies that do not follow the values listed in Table 1 will be left behind.

2. Challenges to face

I believe that it is significant that we were able to publish, with the support of various people, this special issue in 2021, the year in which we began to notice the challenges related to the use of AI "from principles to practice," as stated in the title.

In Japan, it may seem that we are still in the development phase of the principles, or at least in the phase of understanding the meaning and importance of these principles, however discussions in the international community are already a step forward, and are beginning to address the practical application and problems of the principles. In the next section I will briefly introduce some of these.

2.1 Discrepancies between the global and the local conditions

As mentioned in the introduction, the values listed in Table 1 are shared internationally. However, in order to use these values in practice, it is necessary to correctly understand their meaning. For example, the meaning of "fairness," can change depending on whether it refers to the outcome of the process or to the process itself.

Furthermore, a general agreement does not exclude the possibility of a disagreement on some particular issue. In addition, there are some situations in which some countries cannot interfere with the way some values are put into practice. For example, if such countries or regions decide not to accept a particular value, they will be excluded from discussions on that particular value.

As Ms. Gal rightly points out in her article, there are also some concerns regarding the agreement between the decisions and statements of the Japanese international community and the internal discussions. All discussions of values are led by the government, while top-down decisions are made by internal groups of experts, and then disseminated to the international community. Although the international presence is highly appreciated, the sensitivity to the principles of fairness, accountability and transparency (FAT) is very low in domestic companies. A society where cyber and physical technologies are fused, which is the philosophy of Society 5.0 defined by the Japanese Government, is accepted by the international community as a general framework. However, the anthropomorphism of AI and robots, which results in "a society where humans and artifacts live together," is extremely alarming for the international community, especially in Western countries.

How can Japan balance local cultural needs and the need for economic innovation? Ms. Gal, with her profound knowledge of Japanese society, is able to answer this question objectively.

2.2 Exploitation in the name of participation

In order to reduce the risk of discrimination and exclusion in the use of AI, a collaborative approach is considered important right from the development phase. Various methods have been developed to ensure a collaborative approach not only for AI but also for other emerging technologies. After examining these methods, Dr. Park raises the issue that "participation" does not mean inclusion, but can involve exploitation and can be used to justify services and products.

In what Dr. Park describes as the "inclusive illusion," the opinion of the "participants" is often ignored by those in power, and therefore does not influence any decision. The power

gradient is an old but recurring problem, in fact, apparent "participation" is often granted only to those in power, such as AI developers and service providers.

The question is whether we can create a system that not only reaches diverse and marginalized communities, but also allows them to participate in decision-making processes. Such efforts will be especially important when AI will be part of our lives as widely as public institutions and law enforcement agencies.

2.3 Different perceptions among stakeholders

In addition to inclusion, diversity is also important. People have different goals, values, and incentives. However, the more diverse a meeting is, the less coherent the initial discussion will be. I remember that a Japanese bureaucrat once told me frustratedly that he had attended chaotic meetings where participants were very different. He said, "Meetings are held without a direction or agenda and I do not think anything productive can be achieved from such a diverse group of people." This is what a typical Japanese person would say. In fact, for the Japanese society behind-the-scenes negotiations, objectives and agenda approvals are very important and are often already decided before the meeting. However, the people in power should not be the only ones to take decisions.
Ms. Lannquist and her colleagues are also working on global issues such as climate change, slavery, and focusing on the Global South. AI is not just a business tool; it can also be used to solve global problems, and conversely, it could also cause global division.

Responding to global challenges using different point of views requires mutual listening, dialogue, and coordination, which is not easy to achieve. We often attend meetings were the topic is unclear and the discussions, usually incoherent, are always about the same problems. If in the Japanese society, or in the academic field itself, there was a common system of values, it might be possible to make decisions faster and to draw valid conclusions. However, the question now is whether we can move forward with the belief that the pursuit of diversity has benefits for humanity, and whether we can develop human resources capable of supporting such a diverse dialogue.

2.4 AI ethics used as a smokescreen to hide transgressions

The words "AI ethics" and "governance" are often used in this special issue. It is worth mentioning that these terms are not used by the victims of discrimination and harm, but rather by plaintiffs, such as the IT giant companies that cause these problems. Ms. Shah Dand's statement "AI ethics is used as a smokescreen to hide transgressions," should be kept in mind by all those involved in the ethics and governance of AI, including me. Research on AI ethics and governance is often promoted through a collaboration between the industry and academia, and many studies are funded by the industry itself. Nonetheless, as Ms. Shah-Dand points out, often those who suffer from discrimination and prejudice caused by AI and are unable to talk about it, and also those who do talk about it, may still not receive enough support.

As shown in Table 1, the values of AI ethics and governance are mostly shared. However, when they are put into practice, or when their background is decided, they already have inherent biases. In particular, since AI technology deals with issues such as human dignity and equity, it requires creativity to account for various positions, and in particular to decide what is fair. It can be said that an attitude of constant questioning and confirmation of the background is required.

2.5 Balance between cooperation and competition

For whom and for what purpose should AI be developed? Although the discussion tends to be from a business perspective, research is also conducted on the United Nations Sustainable Development Goals (SDGs), contributions to humanity, and solutions to global problems. As these are common goals for humanity, resources must be prevented from being monopolized by one company and their access should not be restricted. Therefore, as pointed out by Ms. Afanasjeva, before investing resources in research and development, it is important to define a strategy that maintains a balance between cooperation and competition.

This is a matter that has been raised repeatedly not only for AI, but also for other cutting-edge technologies. However, a clear solution has not been defined yet.

3. Conclusions

Equity, inclusion, diversity, and cooperation, are values that should be defended and put into practice. In fact, these principles should not simply be shared and accepted in theory, but they should be put into practice as they are an essential prerogative for a successful society. They are difficult concepts to handle, and this is something I realized thanks to this special issue "From Principles to Practice."

In this special issue, we have received reports on the current situation and encouragement from more experienced people who are one or two steps ahead of us. We have no choice but to face and reduce the risks, and to focus on inevitable challenges with determination. The challenge that AI is currently facing is a trans-scientific problem [4] that cannot be solved by science alone. We cannot continue to ignore these questions, and we have no choice but to proceed with the resources we have at our disposal. The road "from principle to practice," is difficult to pursue; however, we need to take a first step, confident that we can count on the help of many people.

[1] https://arxiv.org/abs/1812.04814

[2] http://wilkins.law.harvard.edu/misc/PrincipledAI_FinalGraphic.jpg

[3] For example, Matsumoto and Ema, Policy Recommendation "RCModel, a Risk Chain Model for Risk Reduction in AI Services", Policy Recommendation by the Institute for Future Initiatives, the University of Tokyo, https://ifi.u-tokyo.ac.jp/en/project-news/4828/, (2020)

[4] Alvin, M. Weinberg, Science and Trans-Science, Minerva, vol 10, No.2, p.209-222, 1972.

AI Principles to Practice:
An Introduction to International Activities

April, 2021